The Dissertation Committee for Ying Chen
certifies that this is the approved version of the following dissertation:

Resource Allocation in Service and Logistics Systems

Committee:

John J. Hasenbein, Supervisor

Erhan Kutanoglu, Co-Supervisor

James E. Bickel

Aida Khajavirad

Douglas J. Morrice

Resource Allocation in Service and Logistics Systems

by

Ying Chen, B.E.; M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2016

Dedicated to my parents and my husband.

Acknowledgments

There are many people I wish to thank for making this dissertation possible.

First and foremost, I owe my deepest gratitude to my advisors, Dr. John Hasenbein and Dr. Erhan Kutanoglu, for their insightful guidance and encouragement throughout the process. Both of them have been my mentors in numerous ways, showing me not only the charm of research but also the way of being a better person. They are knowledgeable, patient and supportive, and I have been really fortunate to have them as my advisors.

I would also like to acknowledge Dr. Eric Bickel, Dr. Aida Khajavirad and Dr. Douglas Morrice for being my committee members. Discussions with them have inspired me to better understand my research, and I have benefited from their expertise in different fields especially for technical details.

I am grateful for the various forms of support provided by Dr. David Morton, Dr. Jonathan Bard, Dr. Wesley Barnes, Dr. Nedialko Dimitrov and Dr. Evdokia Nikolova during my graduate study. I am also thankful for the friendship of my fellow students in the Operations Research and Industrial Engineering Program, who are intelligent, enthusiastic and ready to help whenever needed.

Last but not least, my sincere gratitude goes to my dear parents and

my beloved husband, who always have faith in me and encourage me to chase my dream. They are there with love and care, sharing my frustration and happiness. They are my source of strength on each and every day.

Resource Allocation in Service and Logistics Systems

Publication No. _____

Ying Chen, Ph.D.
The University of Texas at Austin, 2016

Supervisor: John J. Hasenbein
Co-Supervisor: Erhan Kutanoglu

Resource allocation is a problem commonly encountered in strategic planning, where a typical objective is to minimize the associated cost or maximize the resulting profit. It is studied analytically and numerically for service and logistics systems in this dissertation, with the major resource being people, services or trucks.

First, a staffing level problem is analyzed for large-scale single-station queueing systems. The system manager operates an Erlang-C queueing system with a quality-of-service (QoS) constraint on the probability that a customer is queued. However, in this model, the arrival rate is uncertain in the sense that even the arrival-rate distribution is not completely known to the manager. Rather, the manager has an estimate of the support of the arrival-rate distribution and the mean. The goal is to determine the number of servers needed to satisfy the quality of service constraint. Two models are explored. First, the constraint is enforced on an overall delay probability, given the

probability that different feasible arrival-rate distributions are selected. In the second case, the constraint has to be satisfied by every possible distribution. For both problems, asymptotically optimal solutions are developed based on Halfin-Whitt type scalings. The work is followed by a discussion on solution uniqueness with a joint QoS constraint and a given arrival-rate distribution in multi-station systems.

Second, an extension to Naor's analysis on the joining or balking problem in observable $M/M/1$ queues and its variant in unobservable $M/M/1$ queues is presented to incorporate parameter uncertainty. The arrival-rate distribution is known to all, but the exact arrival rate is unknown in both cases. The optimal joining strategies are obtained and compared from the perspectives of individual customers, the social optimizer and the profit maximizer, where differences are recognized between the results for systems with deterministic and stochastic arrival rates.

Finally, an integrated ordering and inbound shipping problem is formulated for an assembly plant with a large number of suppliers. The objective is to minimize the annual total cost with a static strategy. Potential transportation modes include full truckload shipping and less than truckload shipping, the former of which allows customized routing while the latter does not. A location-based model is applied in search of near-optimal solutions instead of an exact model with vehicle routing, and numerical experiments are conducted to investigate the insights of the problem.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Resource Allocation

Resource allocation is a crucial concept in strategic planning, where tangible assets such as materials and tools as well as intangible things like services and human resources are distributed over a group of entities. In a broad sense, all questions in this field can be summarized as what to assign and which to assign it to, while the former one is essentially how much to assign if there is only a single resource. The effectiveness and efficiency of answers to the questions are typically evaluated by how well they achieve the managerial objectives, which include but not limited to reducing waiting in a service system, maintaining a smooth material flow in a logistics system, and maybe most important, reducing operational costs or increasing the profit in all cases. This gives a sketch of the problems we consider in this dissertation.

1.2 Service Optimization and Queueing Models

A service system in general refers to a network where service providers such as agents and machines deliver requested services to customers. A more detailed definition can be found on page 11 in [18]. Analyzing and optimizing

such systems have been one of the major subjects in operations research and operations management. In particular, they are usually described by queueing models due to randomness in arrival and service processes. Queueing models have been widely used in telecommunication, traffic engineering, service facility design as well as customer flow control, and there has recently been increased interest in incorporating parameter uncertainty into the models.

In a classical queueing model, customers are assumed to conform to the queueing rules unconditionally. That is to say, they do not make strategic decisions for the sake of their own benefits. We study the staffing problem for large-scale service systems with stochastic arrival rates under this setting, which is motivated by applications in call centers that have demands varying with unpredictable factors. Call centers are labor intensive, and the staffing cost for answering phone calls amounts to about 60-80% of the total operating budget (see [2]). Also, call centers are normally equipped with service quality contracts, so that violations of certain performance constraints, which are likely caused by a lack of representatives, are penalized. We therefore desire to find the lowest staffing level that satisfies the quality requirements. However, obtaining exact values of some quality measures can be computationally challenging when the system is large. In addition, like most optimization problems, optimal staffing of queues requires estimation and prediction of parameters, which include the customer arrival rate (as a characteristic parameter of the arrival distribution itself). Both forecast errors and possible intrinsic randomness of the arrival rate can introduce additional levels of stochasticity that

complicates the analysis of the staffing problem.

An argument about the classical perspective is unlike components waiting to be processed, people waiting in service queues tend to act strategically, which means they attempt maximizing their individual benefits instead of simply following the schedule. Hence, a game-theoretic model is often more realistic for describing service systems involving humans, and research in the area has been blooming in recent decades. We concentrate on the joining or balking problem explored first by Naor [47] in the context of observable queues, where queue lengths are known to arriving customers, as well as its variant in unobservable queues (see [25]), where the information is unknown. Before entering the queue, each customer computes the trade-off between the potential service benefit and the expected waiting cost, and then makes an individual decision on whether to stay for service or leave immediately. The optimal joining strategies from standpoints of individual customers, the social optimizer and the system manager can deviate from each other, and our goal is to coordinate the decisions in systems with random arrival rates by charging appropriate entering fees. A potential application of the work is to price self-ordered medical tests that are not covered by insurance from a marketing view. The problem is classified as a generalized resource allocation problem because we discuss in what scenarios a customer can receive services.

1.3 Inbound Logistics Management

Logistics management is the science that focuses on planning and improving material flow in supply chains. A common goal of research in this field is to optimize the usage of resources, such as transportation equipment and warehouse space, so that the related operating cost is minimized. There are two basic subcategories of logistics activities: inbound logistics and outbound logistics. The former concentrates on purchasing products from suppliers and shipping them to manufacturers, while the latter mainly studies how to stock and send finished goods to consumers. In both cases, freight transportation and inventory holding are potential elements that result in a high operating cost. This happens to the engine assembly plant that suffers from an enormous cost of stocking and inbound shipping and thus motivates our work. The plant is located remotely from most of the contract suppliers, whose products are typically heavy and expensive, due to a recent move from the center of them. Meanwhile, the same ordering and shipping strategy is maintained as before. That is, the suppliers are divided into groups, and each of them is daily served by a predetermined full truckload route to ship parts to the plant. The idea of using daily shipping is to reduce the inventory carrying cost incurred, but it also leads to an unnecessarily high expense of shipping in the current scenario. On the other hand, optimizing the shipping cost alone can cause high inventory levels since low shipping frequencies are preferred in general. We are therefore interested in finding a strategy that decreases the total inbound shipping and inventory holding cost. In particular, we allow the use of less

than truckload shipping here to achieve a more complete analysis.

1.4 Dissertation Organization

The remainder of the dissertation is organized as follows. In Chapter 2, we investigate the static staffing problem in a single-station large-scale service system with an uncertain arrival rate, where we enforce a constraint on the probability of customer delay. We then include a brief discussion on an extension of the problem to a multi-station case in Chapter 3. We take the game-theoretic approach and analyze the balking model with a stochastic arrival rate in Chapter 4. For a logistics system, we formulate a location-based model for an integrated replenishment and inbound shipping problem in Chapter 5, and we develop the insights by numerical experiments. Finally, we summarize the contributions of the dissertation and suggest some future research directions in Chapter 6.

Chapter 2

Staffing Large-Scale Service Systems with Distributional Uncertainty

2.1 Introduction

We investigate the problem of staffing single-station service systems with a quality-of-service (QoS) constraint on the probability of customer delay. We extend previous work in this area by assuming that the arrival rates are uncertain in the sense that even the arrival-rate distribution is unknown. Instead we assume that the support and mean of the distribution have been estimated. The problem is considered from various viewpoints on how "nature" chooses the actual arrival-rate distribution from the feasible set of distributions. In particular, we develop approximately optimal solutions to the staffing problem for large-scale systems under the classical Halfin-Whitt regime. This type of model is useful in service systems operating over a period of months in which, say, some estimates of the peak arrival rate on Monday have been obtained and the QoS constraint is applied to the overall probability that a customer is delayed on a Monday at peak times.

The system structure is depicted in Figure 2.1, where customers arrive according to a Poisson process, wait in the queue when no servers are idle, and

exit from the network after being served. The service times are assumed to be independent and identically distributed exponential random variables, and we normalize the service rate to be 1 without loss of generality. If all the system parameters are known to the decision maker, then the system is an $M/M/s$ queue and the probability of delay is given by the Erlang-C formula. To

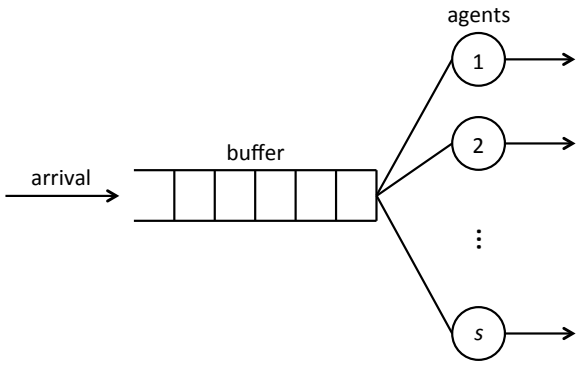


Figure 2.1: Structure of a Single-Station Service System

organize ideas in the chapter, consider the following set of systems categorized by the arrival-rate uncertainties they capture:

- Level I. The arrival rate is a known constant.
- Level II. The arrival rate is a discrete random variable with a known distribution.
- Level III. The arrival rate is a discrete random variable, with known support and mean.

Level I systems have been extensively studied in the classical queueing literature. Level II systems have only been analyzed more recently. This chapter in

particular extends the results of Zan et al. [60]. Although there has been some work on Level III type queueing problems, the particular problem we explore is new to the best of our knowledge. Apart from providing solutions to staffing problems based on the Level III model, we are also interested in the value of information (VOI) when moving from Level III to Level II. In other words, what is the value of knowing the complete arrival-rate distribution, versus knowing just the mean? If the VOI is relatively small, then one implication is that it may not be worth the effort to estimate the entire distribution.

Defining the QoS constraints for the three different models introduced above requires some thought. In a Level I problem, we assume that the constraint is on the probability that an arriving customer has a positive delay, i.e., he does not receive service immediately. In the Level II problem, the constraint is on the expected probability of delay, where the expectation is taken over the known arrival-rate distribution. In the Level III problem, there are several possible choices for a QoS constraint, depending on the decision maker's view of risk and the behavior of nature. We consider two possibilities in this chapter. The first is that among all the feasible arrival-rate distributions, nature chooses a distribution uniformly. Since the set of feasible distributions is defined by a bounded polytope, there is a canonical way to define "uniform" in this case. The other possibility considered is that nature chooses the "worst" distribution, after the decision maker has chosen the number of servers. This is equivalent to the typical adversarial view considered in many models. This view leads to a robust optimization problem whereas the uniform model is a

standard stochastic optimization problem.

In this chapter, we focus on the Level III system, and study two separate models with the different QoS constraints outlined above. The first contribution is to provide asymptotically optimal solutions to the server staffing problem. In addition, we create a method to evaluate the VOI induced by the knowledge differential in the Level II and Level III problems.

The literature on staffing service centers dates back to the foundational work of Erlang at the beginning of the 20^{th} century, including the derivation of the Erlang-C formula. Due to the computational effort involved, various approximations of the service delay probability have been developed. Halfin and Whitt [29] take an asymptotic perspective and provide a limiting result in the so-called quality-and-efficiency-driven (QED) regime. Janssen et al. [36] extend these pioneering results by deriving tight bounds of the service delay probability. For recent studies on optimal staffing in different regimes or with various performance measures, see, for example, Borst et al. [14] and Baron and Milner [7].

With increasing interest in capturing parameter uncertainty, more recent work in service systems has incorporated this idea, usually by assuming the arrival rate is itself random. Chen and Henderson [19] summarize three causes that might result in arrival rates that can be viewed as stochastic: load fluctuation in non-stationary systems, errors of demand forecasting and the inherent randomness of the arrival rate in a doubly stochastic Poisson process.

An important branch of research on staffing service centers with unknown time-varying arrival rates lies in the field of parameter prediction and dynamic staffing by data-driven methods, as seen in Whitt [57], Avramidis et al. [5], Moallemi et al. [45] and Bassamboo and Zeevi [10].

Another body of work focuses on models with given distributions for demand or arrival rate distributions. Harrison and Zeevi [31] balance staffing costs and customer abandonment penalties by solving a newsvendor-type problem for multi-station service centers. Bassamboo et al. [8] then establish a rigorous proof of an asymptotic lower bound on the expected cost implied by the results in [31]. Bassamboo et al. [9] also provide more theoretical results on a newsvendor-type solution for single-station queueing systems. In addition to random arrival rates, Whitt [58] takes uncertain staffing due to absenteeism into consideration. Mandelbaum and Zeltyn [42] solve a constraint satisfaction problem for optimal staffing with different asymptotic regimes. Gans et al. [27] integrate prediction and optimization for staffing call centers. Gurvich et al. [28] formulate a similar problem as a chance-constrained model, and propose a solution approach that can be applied to disparate QoS constraints. Liao et al. [39, 40] discuss the multi-period single-shift staffing plan. Finally, Koçağa et al. [37] allow outsourcing customers when desired. It is also worth noticing that Bandi et al. [6] recently propose an alternative way of modeling queueing systems, where all primitives are described with uncertainty sets rather than renewal processes. However, to the best of our knowledge, there has not been work regarding the lack of perfect distributional information of

the random arrival rate, nor have there been studies on the value of information in such models.

The rest of the chapter is organized as follows. In Section 2.2, we review the theorems on which our analysis is based and illustrate how our methodology works on Level I and Level II problems. Then we discuss a model of the Level III problem and its corresponding solution technique in Section 2.3 and another in Section 2.4. The computational observations are given at the end in Section 2.5.

2.2 Mathematical Background

Solution techniques to lower-level problems are essential for our study on Level III problems. For completeness, we review these background results in this section.

Consider a Level I case where the arrival rate λ is given. Recall that the service rate μ is set to 1 without loss of generality. With a classical $M/M/s$ model, the Erlang-C formula provides the service delay probability $\alpha(s, \lambda)$ for any staffing level $s > \lambda$:

$$\alpha(s, \lambda) := \eta \frac{\lambda^s}{s!(1 - \lambda/s)},$$

where

$$\eta = \left[\sum_{j=0}^{s-1} \frac{\lambda^j}{j!} + \frac{\lambda^s}{s!(1 - \lambda/s)} \right]^{-1}.$$

If $s \leq \lambda$, we define $\alpha(s, \lambda) = 1$ since the system is not stable. Let $c(s)$ denote the staffing cost for a given level s, where $c(\cdot)$ is assumed to be positive

and strictly increasing for $s > 0$. Given a maximum acceptable service delay probability $\epsilon \in (0, 1)$, we define the Level I server staffing problem as follows:

$$\min_s c(s) \qquad \text{s.t.} \quad \alpha(s, \lambda) \leq \epsilon. \tag{2.1}$$

Given our assumptions on the cost function, it is clear that (2.1) reduces to a relatively straightforward root-finding problem.

In the sequel, we make use of a continuous extension of the Erlang-C formula for $\lambda > 0$ and $s \in (\lambda, \infty)$ which appears, for example, in Jagers and Van Doorn [35]:

$$\bar{\alpha}(s, \lambda) := \left[\lambda \int_0^\infty t e^{-\lambda t} (1 + t)^{s-1} dt \right]^{-1}.$$

For $\lambda > 0$ we define $\bar{\alpha}(\lambda, \lambda) = 1$. This corresponds to the situation of an unstable queue, hence it is intuitive to define the delay probability to be 1. It can be shown, using the upper and lower bounds on $\bar{\alpha}$ in Theorem 2.2.2, that $\lim_{s \searrow \lambda} \bar{\alpha}(s, \lambda) = 1$. Note then that for a fixed $\lambda > 0$, $\bar{\alpha}(s, \lambda)$ is right-continuous at λ.

In Theorem 2.2.1, we repeat a classic result of Halfin and Whitt [29]. One implication of the result is that for large arrival rates, an appropriate guide to staffing the system is to use the *square-root safety* staffing rule, i.e., the number of servers should be roughly $\lambda + \beta\sqrt{\lambda}$, where β is the safety staffing factor. If one operates in this regime, then in (2.1) we can view β instead of s as the decision variable.

Theorem 2.2.1. (Halfin and Whitt [29]) *Consider a sequence of $M/M/s$ queues with arrival rates λ_s, $s = 1, 2, \ldots$. As $s \to \infty$, $\alpha(s, \lambda_s)$ converges to a constant α with $0 < \alpha < 1$ if and only if*

$$\sqrt{s}(1 - \rho_s) \to \beta \tag{2.2}$$

for some $\beta > 0$, where $\rho_s = \lambda_s/s$. If (2.2) holds, then

$$\alpha = \frac{1}{1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}}, \tag{2.3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution, and $\phi(\cdot)$ is the corresponding probability density function.

Due to the difficulty in evaluating the Erlang-C formula for large-scale queueing systems, we build our analysis on the square-root safety staffing rule. Theorem 2.2.2 provides upper and lower bounds on $\bar{\alpha}(s, \lambda)$, which are more analytically tractable than the Halfin-Whitt formula. We refer to these henceforth as the JVLZ bounds. The right-hand side of inequality (2.4) is the JVLZ upper bound $UB(\beta, \lambda)$, and we use it to approximate the service delay probability in subsequent sections. By replacing the exact QoS constraint by an upper bound, we guarantee that solutions to the revised problem are feasible in the original problem.

Theorem 2.2.2. (Janssen et al. [36]) *For $\lambda > 0$ and $s > \lambda$, let*

$$\rho = \lambda/s,$$
$$a = \sqrt{-2s(1-\rho+\ln\rho)},$$
$$\beta = (s-\lambda)/\sqrt{\lambda},$$
$$\gamma = (s-\lambda)/\sqrt{s} = \beta\sqrt{\rho}.$$

Then,

$$\bar{\alpha}(s,\lambda) \leq \left[\rho + \gamma\left(\frac{\Phi(a)}{\phi(a)} + \frac{2}{3\sqrt{s}}\right)\right]^{-1}, \tag{2.4}$$

and

$$\bar{\alpha}(s,\lambda) \geq \left[\rho + \gamma\left(\frac{\Phi(a)}{\phi(a)} + \frac{2}{3\sqrt{s}} + \frac{1}{\phi(a)}\frac{1}{12s-1}\right)\right]^{-1}. \tag{2.5}$$

To illustrate the use of the results above, we solve a simple Level I problem.

Example 1. Consider the Level I problem posed in (2.1). Suppose the arrival rate λ is 400 calls per minute, the average call time is 1 minute, and the service level threshold value $\epsilon = 0.30$.

To solve this problem by Theorem 2.2.1, we invert equation (2.3) with $\alpha = \epsilon$, which yields a safety staffing factor of $\beta = 0.829$. The staffing rule then indicates that the staffing level should be set to $\lceil (400 + 0.829 \cdot \sqrt{400}) \rceil = 417$.

In fact, it can be checked that this approximate solution indeed is optimal in the original problem. One way to do so is to compute the delay probabilities exactly, which is not an easy task due to numerical instability of the Erlang-C formula. Instead, we can use the JVLZ bounds. When $s = 417$, the

right-hand side of equation (2.4) is 0.297, so the solution to problem (2.1) is at most 417. On the other hand, the JVLZ lower bound obtained from equation (2.5) is 0.322 when $s = 416$, which indicates the optimality of $s = 417$. $\square$

We now consider the Level II model in which the decision maker does not know the arrival rate precisely, but rather knows only the distribution of the rate. We denote the arrival rate with a random variable Λ, which has a discrete state space $\Omega = \{\lambda^{\omega_1}, \ldots, \lambda^{\omega_n}\}$ with $\mathbb{P}\{\Lambda = \lambda^{\omega_k}\} = p^{\omega_k}$ for $k \in \{1, \ldots, n\}$. Assume $\lambda^{\omega_k} < \lambda^{\omega_l}$ for $k < l$ without loss of generality.

The Level II staffing problem is as follows:

$$\min_s c(s) \qquad \text{s.t.} \quad \mathbb{E}_\Lambda[\bar{\alpha}(s, \Lambda)] \leq \epsilon. \tag{2.6}$$

If we imagine a service system being staffed over a sequence of days, each of which sees a realization of Λ, then the constraint in (2.6) can be viewed as requiring that the proportion of customers experiencing a delay, when averaged over days, is no more than ϵ.

For large values in the state space Ω we seek to simplify the computation by applying the square-root safety staffing rule and the JVLZ approximation. Zan et al. [60] outline a methodology to solve (2.6), which we now summarize. The first step is to define a *key* scenario ω^{key} which serves as a base scenario. Once the key scenario is chosen, we perform a "micro-optimization" by tuning β based on the key scenario.

There can be multiple base scenarios that lead to the optimal staffing level since the safety factor can take on any positive value. We wish to define

the key scenario as the largest feasible base scenario. For any feasible s in problem (2.6), we define

$$\hat{\omega}^{key}(s) := \max_{k\in\{1,\ldots,n\}} \{\omega_k \mid s \geq \lambda^{\omega_k}\}. \tag{2.7}$$

With this definition, the scenario associated with the optimal s must be unique, and so is the β.

To attack the Level II problem we take an asymptotic view of problem (2.6). That is, we imagine the service system is gradually expanded in the following way. Assume there is a sequence of random variables $\Lambda_1, \Lambda_2, \ldots$ denoting the growing arrival rates. The possible realizations of Λ_m for $m \in \mathbb{Z}^+$ belong to $\{\lambda_m^{\omega_1}, \ldots, \lambda_m^{\omega_n}\}$ with $\lambda_m^{\omega_k} = m\lambda_1^{\omega_k}$, so the arrival rate tends to infinity as m increases. Under such a scaling, the "gaps" between scenarios grow large so that in the limit the probability of having customers waiting is 0 for scenarios with lower arrival rates than $\lambda^{\omega^{key}}$, and 1 for those with higher scenarios, if $\lambda^{\omega^{key}}$ is used to set the staffing level.

We write the problem with any given m as

$$\min_s c(s) \qquad \text{s.t.} \quad \mathbb{E}_{\Lambda_m}[\bar{\alpha}(s, \Lambda_m)] \leq \epsilon. \tag{2.8}$$

For $0 < \epsilon < 1$, Zan et al. [60] identified a scenario can serve as the basis for an optimal solution to problem (2.8) for some $\beta \geq 0$ when m is sufficiently large:

$$\omega^{key}(\mathbf{p}) := \begin{cases} \omega_i, & \text{if } \sum_{k=i}^{n} p^{\omega_k} \geq \epsilon \text{ and } \sum_{k=i+1}^{n} p^{\omega_k} < \epsilon, \\ & \forall i \in \{1, \ldots, n-1\}; \\ \omega_n, & \text{otherwise.} \end{cases} \tag{2.9}$$

Notice that $\omega^{key}(\mathbf{p})$ depends on nothing but $\mathbf{p}$, where $\mathbf{p} = (p^{\omega_1}, \ldots, p^{\omega_n})^T$. In other words, the optimal staffing level can always be attained by applying the square-root safety rule with the arrival rate $\lambda_m^{\omega^{key}(\mathbf{p})}$ and some appropriately chosen nonnegative safety factor. Theorem 16 from Zan et al. [60] has formalized this idea, which we will restate in Theorem 2.2.4 using our definition of $\hat{\omega}^{key}$ from equation (2.7), with Lemma 2.2.3 ensuring the validity of our translation.

Lemma 2.2.3. *For $m \in \mathbb{Z}^+$ let $s_m^\star$ be the solution to (2.8). Then there exists an $\bar{m}$ such that for all $m \geq \bar{m}$, $\hat{\omega}^{key}(s_m^\star) = \omega^{key}(\mathbf{p})$.*

Proof. Assume $\omega^{key}(\mathbf{p}) = \omega_l$. Zan et al. [60] proved that there exists an $\tilde{m}$ such that for all $m \geq \tilde{m}$, $s_m^\star \geq \lambda_m^{\omega_l}$.

Let $\Delta = \epsilon - \sum_{k=l+1}^{n} p^{\omega_k}$. Note that Δ must be positive due to the definition of $\omega^{key}(\mathbf{p})$. For each k such that $p^{\omega_k} \neq 0$, set $\delta_k = \frac{\Delta}{lp^{\omega_k}}$. We define $\tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) = \bar{\alpha}(\lambda_m^{\omega_l} + \beta\sqrt{\lambda_m^{\omega_l}}, \lambda_m^{\omega_k})$. By Corollary 13 in [60], for any $k \leq l$ and $\beta \geq 0$

$$\lim_{m\to\infty} \tilde{\alpha}(\beta, \lambda_m^{\omega_{l+1}}, \lambda_m^{\omega_k}) = 0.$$

Hence, for each δ_k there exists an $\bar{m}_k$ such that $\tilde{\alpha}(\beta, \lambda_m^{\omega_{l+1}}, \lambda_m^{\omega_k}) < \delta_k$ for all $m \geq \bar{m}_k$. Let $\bar{m}$ be the larger of $\tilde{m}$ and the maximum of these $\bar{m}_k$ values. We claim that $\lambda_m^{\omega_{l+1}} > s_m^\star$ for all $m \geq \bar{m}$. Suppose there exists an $\hat{m} \geq \bar{m}$ for which $\lambda_{\hat{m}}^{\omega_{l+1}} \leq s_{\hat{m}}^\star$. Then, problem (2.8) can be rewritten as:

$$\min_{\beta\geq 0} \bar{c}(\beta, \omega_{l+1}) \qquad \text{s.t.} \quad \sum_{k=1}^{n} p^{\omega_k} \tilde{\alpha}(\beta, \lambda_m^{\omega_{l+1}}, \lambda_m^{\omega_k}) \leq \epsilon, \tag{2.10}$$

where $\bar{c}(\beta, \omega_{l+1}) = c(\lambda_m^{\omega_{l+1}} + \beta\sqrt{\lambda_m^{\omega_{l+1}}})$. Then by construction, for this $\hat{m}$ the constraint in (2.10) is not active, which contradicts the optimality of $s_{\hat{m}}^\star$. We have now shown that $\lambda_m^{\omega_{l+1}} > s_m^\star \geq \lambda_m^{\omega_l}$, establishing the lemma. □

The logic in the proof of Lemma 2.2.3 implies that (2.8) can be written as:

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad \text{s.t.} \quad \sum_{k=1}^{n} p^{\omega_k} \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \leq \epsilon, \tag{2.11}$$

recalling that we defined $\omega_l = \omega^{key}(\mathbf{p})$. Furthermore, if m is sufficiently large, then we can approximate $\tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k})$ by $UB(\beta, \lambda_m^{\omega_l})$ for $k = l$, by 0 if $k < l$, and by 1 otherwise. This suggests that we can form the following approximate version of (2.11):

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad \text{s.t.} \quad p^{\omega_l} UB(\beta, \lambda_m^{\omega_l}) \leq \left(\epsilon - \sum_{k=l+1}^{n} p^{\omega_k} \right). \tag{2.12}$$

The following result is a modification of Theorem 16 in [60] which connects the exact Level II problem in (2.8) and the approximate version in (2.12).

Theorem 2.2.4. (Zan et al. [60]) *Fix $\epsilon \in (0, 1)$. Let $s_m^\star$ be the solution to* (2.8) *and β_m^G be an optimal solution to model* (2.12) *for $m \in \mathbb{Z}^+$. Then, there exists an $\bar{m}$ such that for all $m \geq \bar{m}$, $\omega^{key}(\mathbf{p}) = \hat{\omega}^{key}(s_m^\star)$. And, there exists a $\beta^\star \geq 0$ such that*

$$\lim_{m \to \infty} \beta_m^G = \lim_{m \to \infty} \beta_m^F = \beta^\star,$$

where the β_m^F are the optimal staffing factors for (2.11) *for all m.*

Theorem 2.2.4 justifies the validity of approximating the optimal solution to problem (2.8) by identifying $\omega^{key}(\mathbf{p})$ and solving problem (2.12) instead when m is large. Example 2 demonstrates how we can solve a simple large-scale Level II problem by this procedure.

Example 2. Consider the Level II problem (2.6) with $\Omega = \{100, 200, 400\}$, $\mathbf{p} = (0.58, 0.38, 0.04)$ and $\epsilon = 0.30$.

From (2.9) the key scenario ω_2 with $\lambda^{\omega_2} = 200$. Hence we need the delay probability requirement in ω_2 to be such that $0.04 + 0.38 \cdot UB(\beta, \lambda^{\omega_2}) = 0.30$, which gives us $UB(\beta, \lambda^{\omega_2}) = 0.684$. We get $\beta = 0.294$ by inverting the JVLZ upper bound. The approximate solution to (2.6) is then given by $\lceil (200 + 0.294 \cdot \sqrt{200}) \rceil = 205$. It can be checked that in fact the optimal solution to the original problem is also 205. □

With Level I and Level II problems reviewed, we are prepared to solve Level III problems. Suppose the system controller does not know the exact value of $\mathbf{p}$, but instead knows only Ω and that $E[\Lambda] = r$. Note that r must be such that $\lambda^{\omega_1} < r < \lambda^{\omega_n}$. How should we staff a service system like this? We address this question in the next sections.

2.3 Level III Models with Meta-Distributions

In this section, we analyze the Level III model and a related optimization problem. We assume that the decision maker knows Ω and also knows $E[\Lambda] = r$. Since Ω is finite, the set of possible discrete distributions with a

fixed mean is defined by the following polytope:

$$\sum_{k=1}^{n} p^{\omega_k} \lambda^{\omega_k} = r, \tag{2.13}$$

$$\sum_{k=1}^{n} p^{\omega_k} = 1, \tag{2.14}$$

$$p^{\omega_k} \geq 0, \forall k \in \{1, \ldots, n\}. \tag{2.15}$$

Let $\mathcal{D}$ denote the set of vectors $\mathbf{p}$ satisfying equations (2.13)-(2.15). Each element in $\mathcal{D}$ then corresponds to a potential discrete arrival-rate distribution. We assume nature picks a vector from $\mathcal{D}$ following a particular distribution. In this case, we call the random object with state space $\mathcal{D}$ a meta-random variable D.

For any such D, one Level III optimization problem is defined as follows:

$$\min_{s} c(s) \qquad \text{s.t.} \quad \mathbb{E}_D \left\{ \mathbb{E}_{\Lambda(D)}[\bar{\alpha}(s, \Lambda(D))] \right\} \leq \epsilon. \tag{2.16}$$

Although this problem is more complex than the Level II problem introduced in the previous section, note that structurally it is still a one-stage stochastic optimization problem. We first provide some results for a generally distributed D and then discuss the special case when D has a uniform distribution.

As before, we are interested in systems parameterized by $m \in \mathbb{Z}^+$ in which the m grows large. The asymptotic optimality result below directly extends Theorem 2.2.4 to the Level III case. In a sense, the result implies that the Level III problem can be first reduced to a Level II problem, which can then be solved by the methods in the previous section.

We next wish to define a sequence of Level III problems with arrival rates that grow with m. Consider first a base problem with state space Ω and $E[\Lambda] = r$. These quantities define the base polytope $\mathcal{D}$. In scaled versions of the problem, the polytope remains the same, but the state space of the arrival rate is given by $\Omega_m = \{m\lambda^{\omega_1}, \ldots, m\lambda^{\omega_n}\}$, and we denote the associated scaled random arrival rate with Λ_m or $\Lambda_m(D)$ if we want to emphasize the dependence on the value of D. Notice then that for each D, $E[\Lambda(D)] = rm$ as desired. We are now prepared to present the primary result of this section.

Theorem 2.3.1. *Let* $\mathbf{p}_{\text{M}} := \mathbb{E}_D[\mathbf{p}(D)]$. *Fix* $\epsilon \in (0,1)$. *Let* $s_m^\star$ *be the optimal solution to*

$$\min_s c(s) \qquad s.t. \quad \mathbb{E}_D\left\{\mathbb{E}_{\Lambda_m(D)}\left[\bar{\alpha}(s, \Lambda_m(D))\right]\right\} \leq \epsilon, \tag{2.17}$$

for $m \in \mathbb{Z}^+$. *Set* $\omega_l = \omega^{key}(\mathbf{p}_{\text{M}})$. *Let* β_m^G *be an optimal solution to the model*

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad s.t. \quad p_{\text{M}}^{\omega_l} UB(\beta, \lambda_m^{\omega_l}) \leq \left(\epsilon - \sum_{k=l+1}^{n} p_{\text{M}}^{\omega_k}\right).$$

Then, there exists an $\bar{m}$ *such that for all* $m \geq \bar{m}$, $\omega_l = \hat{\omega}^{key}(s_m^\star)$. *And, there exists a* $\beta^\star \geq 0$ *such that*

$$\lim_{m\to\infty} \beta_m^G = \lim_{m\to\infty} \beta_m^F = \beta^\star,$$

where for all m *the* β_m^F *are the optimal staffing factors for the problem*

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad s.t. \quad \sum_{k=1}^{n} p_{\text{M}}^{\omega_k} \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \leq \epsilon.$$

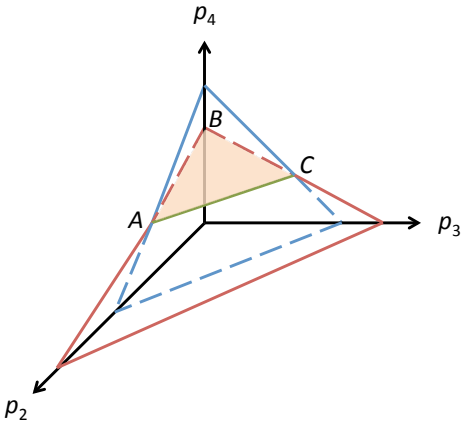
Proof. Consider the left-hand side of the constraint in problem (2.17):

$$\begin{aligned}
\mathbb{E}_D\left\{\mathbb{E}_{\Lambda_m(D)}\left[\bar{\alpha}(s,\Lambda_m(D)\right]\right\} &= \mathbb{E}_D\left[\sum_{k=1}^{n} p^{\omega_k}(D)\bar{\alpha}(s,\lambda_m^{\omega_k})\right] \\
&= \sum_{k=1}^{n}\mathbb{E}_D\left[p^{\omega_k}(D)\bar{\alpha}(s,\lambda_m^{\omega_k})\right] \\
&= \sum_{k=1}^{n}\bar{\alpha}(s,\lambda_m^{\omega_k})\mathbb{E}_D[p^{\omega_k}(D)] \\
&= \sum_{k=1}^{n} p_{\mathrm{M}}^{\omega_k}\bar{\alpha}(s,\lambda_m^{\omega_k}).
\end{aligned}$$

That is, the problem in (2.17) can be viewed as a Level II problem where Λ has a probability mass function given by $\mathbf{p}_{\mathrm{M}}$. Then, the claims in the theorem follow immediately from Theorem 2.2.4. □

We now analyze the special case in which nature is apathetic, i.e., D is a uniform random variable on $\mathcal{D}$. We call this the uniformly meta-distributed (UMD) model. In this case, $\mathbf{p}_{\mathrm{M}}$ coincides with the arithmetic mean, or centroid, of $\mathcal{D}$. Rademacher [50] proved that in general "it is #P-hard to compute the centroid of an [sic] polytope given as an intersection of halfspaces" exactly. Nonetheless, we can compute the centroid analytically for $n \leq 4$. We only show the solution for $n = 4$ since the calculations are straightforward for $n \leq 3$.

When $n = 4$, $\mathcal{D}$ can be described equivalently by equations (2.18)-



$$A = \left(\frac{\lambda^{\omega_4} - r}{\lambda^{\omega_4} - \lambda^{\omega_2}}, 0, \frac{r - \lambda^{\omega_2}}{\lambda^{\omega_4} - \lambda^{\omega_2}} \right)$$
$$B = \left(0, 0, \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_4} - \lambda^{\omega_1}} \right)$$
$$C = \left(0, \frac{\lambda^{\omega_4} - r}{\lambda^{\omega_4} - \lambda^{\omega_3}}, \frac{r - \lambda^{\omega_3}}{\lambda^{\omega_4} - \lambda^{\omega_3}} \right)$$
$$\text{Centroid} = (A + B + C)/3$$

Figure 2.2: Projection of $\mathcal{D}$ when $\lambda^{\omega_3} \leq r < \lambda^{\omega_4}$

(2.21):

$$(\lambda^{\omega_2} - \lambda^{\omega_1})p^{\omega_2} + (\lambda^{\omega_3} - \lambda^{\omega_1})p^{\omega_3} + (\lambda^{\omega_4} - \lambda^{\omega_1})p^{\omega_4} = r - \lambda^{\omega_1}, \tag{2.18}$$

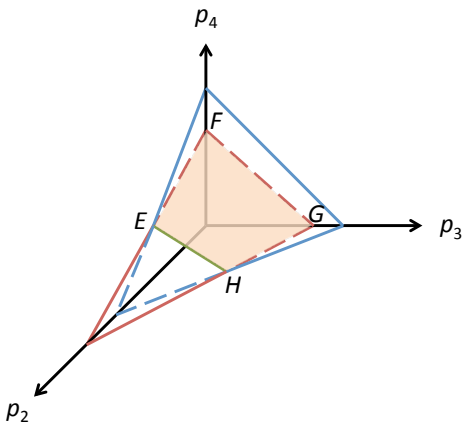
$$p^{\omega_2} + p^{\omega_3} + p^{\omega_4} \leq 1, \tag{2.19}$$

$$p^{\omega_k} \geq 0, \forall k \in \{2, 3, 4\}, \tag{2.20}$$

$$p^{\omega_1} = 1 - p^{\omega_2} - p^{\omega_3} - p^{\omega_4}. \tag{2.21}$$

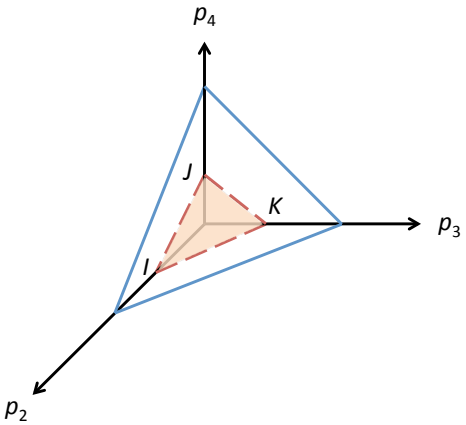
Hence, we can take the projection of $\mathcal{D}$ on the space $(p^{\omega_2}, p^{\omega_3}, p^{\omega_4})$, find the centroid of the two-dimensional figure defined by constraints (2.18)-(2.20), and eventually calculate p^{ω_1} using equation (2.21). The resulting vector $\mathbf{p}_{\text{U}}$ corresponds to $\mathbf{p}_{\text{M}}$ for the UMD model. The projection is illustrated by the shaded areas in Figures 2.2 through 2.4, for various cases. The algorithms for computing $(p_{\text{U}}^{\omega_2}, p_{\text{U}}^{\omega_3}, p_{\text{U}}^{\omega_4})$ are listed as well.

When $n > 4$, we can approximate the centroid with Monte Carlo techniques such as the one implemented in Section 2.5. We complete the discussion



$$E = \left(\frac{\lambda^{\omega_4} - r}{\lambda^{\omega_4} - \lambda^{\omega_2}}, 0, \frac{r - \lambda^{\omega_2}}{\lambda^{\omega_4} - \lambda^{\omega_2}}\right)$$
$$F = \left(0, 0, \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_4} - \lambda^{\omega_1}}\right)$$
$$G = \left(0, \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_3} - \lambda^{\omega_1}}, 0\right)$$
$$H = \left(\frac{\lambda^{\omega_3} - r}{\lambda^{\omega_3} - \lambda^{\omega_2}}, \frac{r - \lambda^{\omega_2}}{\lambda^{\omega_3} - \lambda^{\omega_2}}, 0\right)$$
$$O_1 = (E + F + G)/3$$
$$O_2 = (F + G + H)/3$$
$$O_3 = (G + H + E)/3$$
$$O_4 = (H + E + F)/3$$
Centroid = Intersection of segments O_1O_3 and O_2O_4

Figure 2.3: Projection of $\mathcal{D}$ when $\lambda^{\omega_2} \leq r < \lambda^{\omega_3}$



$$I = \left(\frac{r - \lambda^{\omega_1}}{\lambda^{\omega_2} - \lambda^{\omega_1}}, 0, 0\right)$$
$$J = \left(0, 0, \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_4} - \lambda^{\omega_1}}\right)$$
$$K = \left(0, \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_3} - \lambda^{\omega_1}}, 0\right)$$
Centroid $= (I + J + K)/3$

Figure 2.4: Projection of $\mathcal{D}$ when $\lambda^{\omega_1} \leq r < \lambda^{\omega_2}$

in this section by applying the UMD model to a Level III problem with $n = 4$.

Example 3. Let $\Omega = \{100, 200, 400, 700\}$ with $r = 250$ in problem (2.16). Set the QoS requirement to $\epsilon = 0.30$.

Given that the cardinality of the state space is four, we compute the centroid of $\mathcal{D}$ analytically rather than approximate it. According to Figure 2.3, the centroid is located at the point $\mathbf{p}_{\text{U}} = (0.3542, 0.3625, 0.1875, 0.0958)$. We then follow the steps in Example 2 to solve a Level II problem with the arrival-rate distribution defined by the centroid.

Knowing that the optimal key scenario has arrival rate $\lambda^{\omega_2} = 200$, we get $UB(\beta, \lambda^{\omega_2}) = 0.046$ by solving $0.1875 + 0.0958 + 0.3625 \cdot UB(\beta, \lambda^{\omega_2}) = 0.30$. We then invert $UB(\beta, \lambda^{\omega_2})$, and use $\beta = 1.830$ in the square-root safety staffing rule. This yields a solution to the UMD model of $\lceil (200 + 1.830 \cdot \sqrt{200}) \rceil = 226$. □

2.4 Robust Analysis

A potential consequence of modeling the problem as described in Section 2.3 is that the QoS constraint may not be satisfied by the "true" arrival-rate distribution. In this section, we take a more conservative view by asking how many servers are needed such that the bound on the expected service delay probability is satisfied even when the "worst" distribution occurs. Let d denote a particular element in $\mathcal{D}$, which corresponds to a discrete distribution. With notation for all other expressions carried over from previous sections, we

formulate the robust staffing problem as follows:

$$\min_{s} c(s) \qquad \text{s.t.} \quad \max_{d\in\mathcal{D}} \left\{\mathbb{E}_{\Lambda(d)}[\bar{\alpha}(s,\Lambda(d))]\right\} \leq \epsilon. \tag{2.22}$$

We again use the square-root safety staffing rule, approximate the service delay probability by the JVLZ upper bound, and solve the large-scale robust problem from the asymptotic perspective introduced in Section 2.2. We provide the results in Theorem 2.4.2 and Theorem 2.4.4 after we develop a building block for both of the proofs, Lemma 2.4.1.

Lemma 2.4.1. *For a fixed $k' \in \{1,\ldots,n\}$ and $\beta \geq 0$, let*

$$\mathbf{p}_{\mathrm{T}} \in \arg\max_{d\in\mathcal{D}} \left\{ p^{\omega_{k'}}(d)\tilde{\alpha}(\beta,\lambda^{\omega_{k'}},\lambda^{\omega_{k'}}) + \sum_{k=k'+1}^{n} p^{\omega_k}(d) \right\}. \tag{2.23}$$

Then for any $k \in \{1,\ldots,n\} \setminus \{1,k',k'+1\}$, $p_{\mathrm{T}}^{\omega_k} = 0$.

Proof. We prove the statement by contradiction. Suppose there exists a distribution $\mathbf{p}_{\mathrm{T}}$ which is a maximizer of (2.23) and for which $p_{\mathrm{T}}^{\omega_k} > 0$ for $k \notin \{1,k',k'+1\}$. If $k \in \{2,\ldots,k'-1\}$, we define a new distribution $\tilde{\mathbf{p}}$ with

$$\tilde{p}^{\omega_1} = p_{\mathrm{T}}^{\omega_1} + \frac{\lambda^{\omega_{k'}} - \lambda^{\omega_k}}{\lambda^{\omega_{k'}} - \lambda^{\omega_1}} p_{\mathrm{T}}^{\omega_k},$$

$$\tilde{p}^{\omega_{k'}} = p_{\mathrm{T}}^{\omega_{k'}} + \frac{\lambda^{\omega_k} - \lambda^{\omega_1}}{\lambda^{\omega_{k'}} - \lambda^{\omega_1}} p_{\mathrm{T}}^{\omega_k},$$

$\tilde{p}^{\omega_k} = 0$ and equal probabilities for all other scenarios in $\mathbf{p}_{\mathrm{T}}$. Since $\frac{\lambda^{\omega_k}-\lambda^{\omega_1}}{\lambda^{\omega_{k'}}-\lambda^{\omega_1}}$ and $\tilde{\alpha}(\beta,\lambda^{\omega_{k'}},\lambda^{\omega_{k'}})$ are both positive, $\tilde{\mathbf{p}}$ returns a larger value of the argument in (2.23) than $\mathbf{p}_{\mathrm{T}}$, leading to a contradiction. If $k \in \{k'+2,\ldots,n\}$, define $\tilde{\mathbf{p}}$ as

$$\tilde{p}^{\omega_1} = p_{\mathrm{T}}^{\omega_1} + \frac{\lambda^{\omega_{k'+1}} - \lambda^{\omega_k}}{\lambda^{\omega_{k'+1}} - \lambda^{\omega_1}} p_{\mathrm{T}}^{\omega_k},$$

$$\tilde{p}^{\omega_{k'+1}} = p_{\mathrm{T}}^{\omega_{k'+1}} + \frac{\lambda^{\omega_k} - \lambda^{\omega_1}}{\lambda^{\omega_{k'+1}} - \lambda^{\omega_1}} p_{\mathrm{T}}^{\omega_k},$$

$\tilde{p}^{\omega_k} = 0$ and keep all other probabilities unchanged from $\mathbf{p}_{\mathrm{T}}$. A contradiction is again obtained since $\frac{\lambda^{\omega_k} - \lambda^{\omega_1}}{\lambda^{\omega_{k'+1}} - \lambda^{\omega_1}} > 1$. □

Theorem 2.4.2. *Fix $\epsilon \in (0,1)$ and let ω_l be the corresponding scenario, as defined by Table 2.1. Let $s_m^\star$ be the optimal solution to*

$$\min_s c(s) \qquad s.t. \quad \max_{d \in \mathcal{D}} \left\{ \mathbb{E}_{\Lambda_m(d)}[\bar{\alpha}(s, \Lambda_m(d))] \right\} \leq \epsilon, \tag{2.24}$$

for $m \in \mathbb{Z}^+$. Then, there exists an $\bar{m}$ such that for all $m \geq \bar{m}$, $\omega_l = \hat{\omega}^{key}(s_m^\star)$.

Table 2.1: Solution to the Robust Problem

ϵ	ω_l	$p_{\mathrm{R}}^{\omega_l}$	ν
$\left(0, \frac{r-\lambda^{\omega_1}}{\lambda^{\omega_n}-\lambda^{\omega_1}}\right]$	ω_n	$\frac{r-\lambda^{\omega_1}}{\lambda^{\omega_n}-\lambda^{\omega_1}}$	ϵ
$\left(\frac{r-\lambda^{\omega_1}}{\lambda^{\omega_{i+1}}-\lambda^{\omega_1}}, \frac{r-\lambda^{\omega_1}}{\lambda^{\omega_i}-\lambda^{\omega_1}}\right]$	ω_i	$\min\left(\frac{\lambda^{\omega_{i+1}}-r}{\lambda^{\omega_{i+1}}-\lambda^{\omega_i}}, \frac{r-\lambda^{\omega_1}}{\lambda^{\omega_i}-\lambda^{\omega_1}}\right)$	$\min\left(\epsilon - \frac{r-\lambda^{\omega_i}}{\lambda^{\omega_{i+1}}-\lambda^{\omega_i}}, \epsilon\right)$
$\left(\frac{r-\lambda^{\omega_1}}{\lambda^{\omega_2}-\lambda^{\omega_1}}, 1\right)$	ω_1	$\frac{\lambda^{\omega_2}-r}{\lambda^{\omega_2}-\lambda^{\omega_1}}$	$\epsilon - \frac{r-\lambda^{\omega_1}}{\lambda^{\omega_2}-\lambda^{\omega_1}}$

Proof. We consider a partition of the interval $(0,1]$:

$$\mathcal{Q} = \{\mathcal{Q}_{k'} : k' \in \{1, \ldots, n\}\},$$

where

$$\mathcal{Q}_{k'} = \left(\max_{d \in \mathcal{D}} \left\{ \sum_{k=k'+1}^{n} p^{\omega_k}(d) \right\}, \max_{d \in \mathcal{D}} \left\{ \sum_{k=k'}^{n} p^{\omega_k}(d) \right\} \right].$$

Note that the lower limit of $\mathcal{Q}_n$ is 0 by definition. For any given $\epsilon \in (0,1)$, let $\omega_l = \omega_{k'}$ if $\epsilon \in \mathcal{Q}_{k'}$. We notice that

$$\omega_l = \max_{d \in \mathcal{D}} \omega^{key}(\mathbf{p}(d)).$$

Due to the compactness of $\mathcal{D}$, according to Lemma 2.2.3, there must be an $\bar{m}$ such that for all $m \geq \bar{m}$,

$$\hat{\omega}^{key}(s_m^\star) = \max_{d \in \mathcal{D}} \omega^{key}(\mathbf{p}(d)),$$

and thus $\omega_l = \hat{\omega}^{key}(s_m^\star)$ for $m \geq \bar{m}$.

To compute the interval $\mathcal{Q}_{k'}$, we first consider the optimization problem in the upper limit, which is exactly problem (2.23) with $\beta = 0$. Using Lemma 2.4.1, we simplify the upper limit to

$$\begin{aligned}
\max_{p^{\omega_1}, p^{\omega_{k'}}, p^{\omega_{k'+1}}} \quad & p^{\omega_{k'}} + p^{\omega_{k'+1}} && \text{(2.25a)} \\
\text{s.t.} \quad & p^{\omega_1}\lambda^{\omega_1} + p^{\omega_{k'}}\lambda^{\omega_{k'}} + p^{\omega_{k'+1}}\lambda^{\omega_{k'+1}} = r && \text{(2.25b)} \\
& p^{\omega_1} + p^{\omega_{k'}} + p^{\omega_{k'+1}} = 1 && \text{(2.25c)} \\
& p^{\omega_k} \geq 0, \forall k \in \{1, k', k'+1\}, && \text{(2.25d)}
\end{aligned}$$

where p^{ω_k} is defined to be 0 for $k > n$. Problem (2.25) can be easily solved by variable substitution. The objective value is $\frac{r-\lambda^{\omega_1}}{\lambda^{\omega_{k'}}-\lambda^{\omega_1}}$ if $r \leq \lambda^{\omega_{k'}}$, and 1 otherwise. We directly apply this result to the lower limit by replacing k' with $k'+1$. Finally, the results of this analysis appear in the first two columns of Table 2.1. □

Notice that some of the intervals in Table 2.1 may be empty for certain parameter combinations. For example, we observe that $\omega_l \neq \omega_i$ if $\lambda^{\omega_{i+1}} \leq r$.

Now we can rewrite model (2.24) equivalently as

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad \text{s.t.} \quad \max_{d \in \mathcal{D}} \left\{ \sum_{k=1}^{n} p^{\omega_k}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \right\} \leq \epsilon \tag{2.26}$$

when m is sufficiently large. We focus on the left-hand side of the constraint first. Rather than solving for the worst distribution directly, we consider a natural approximate problem that is easier to handle:

$$\max_{d \in \mathcal{D}} \left\{ p^{\omega_l}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}(d) \right\}. \tag{2.27}$$

Lemma 2.4.3. *For any given $\beta \geq 0$,*

$$\lim_{m \to \infty} \max_{d \in \mathcal{D}} \left| \sum_{k=1}^{n} p^{\omega_k}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) - \left(p^{\omega_l}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}(d) \right) \right| = 0, \tag{2.28}$$

and

$$\lim_{m \to \infty} \max_{d \in \mathcal{D}} \left\{ \sum_{k=1}^{n} p^{\omega_k}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \right\} = \lim_{m \to \infty} \max_{d \in \mathcal{D}} \left\{ p^{\omega_l}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}(d) \right\}. \tag{2.29}$$

Proof. We prove (2.28) first. Assume d_m is the optimal solution to the maximization problem inside the limit for each $m \in \mathbb{Z}^+$. Then we can rewrite the left-hand side of (2.28) as

$$\lim_{m \to \infty} \left| \sum_{k=1}^{l-1} p^{\omega_k}(d_m) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) + \sum_{k=l+1}^{n} p^{\omega_k}(d_m) \left(\tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) - 1 \right) \right|. \tag{2.30}$$

Using the triangle inequality, we derive an upper bound for (2.30):

$$\lim_{m\to\infty}\left[\sum_{k=1}^{l-1} p^{\omega_k}(d_m)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k}) + \sum_{k=l+1}^{n} p^{\omega_k}(d_m)\left(1-\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k})\right)\right]. \quad (2.31)$$

As m goes to infinity, $\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k})$ converges pointwise to 0 when $k<l$ and to 1 when $k>l$ (see Corollary 13 from Zan et al. [60]). Using this observation and the fact that $p^{\omega_k}(d_m)$ is uniformly bounded above by 1 for all m, we conclude that the limit in (2.31) is 0, establishing (2.28).

We now consider (2.29). We define the following quantities

$$d_m^F \in \arg\max_{d\in\mathcal{D}}\left\{\sum_{k=1}^{n} p^{\omega_k}(d)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k})\right\}$$

and

$$d_m^H \in \arg\max_{d\in\mathcal{D}}\left\{p^{\omega_l}(d)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}(d)\right\}.$$

Then, the result in (2.28) indicates that for any $\delta>0$, there exists an $\bar{m}$ such that for all $m>\bar{m}$,

$$\begin{aligned}
&\left|\sum_{k=1}^{n} p^{\omega_k}\left(d_m^F\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k}) - \left(p^{\omega_l}\left(d_m^H\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}\left(d_m^H\right)\right)\right| \\
&\leq \max\left\{\sum_{k=1}^{n} p^{\omega_k}\left(d_m^F\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k}) - \left(p^{\omega_l}\left(d_m^F\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}\left(d_m^F\right)\right),\right. \\
&\qquad \left.\left(p^{\omega_l}\left(d_m^H\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_l}) + \sum_{k=l+1}^{n} p^{\omega_k}\left(d_m^H\right)\right) - \sum_{k=1}^{n} p^{\omega_k}\left(d_m^H\right)\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k})\right\} \\
&\leq \delta.
\end{aligned}$$

This establishes (2.29). □

Theorem 2.4.4. *Fix $\epsilon \in (0,1)$ and let ω_l, $p_{\mathrm{R}}^{\omega_l}$ and ν be the corresponding quantities, as defined by Table 2.1. Let β_m^G be an optimal solution to the model*

$$\min_{\beta \geq 0} \bar{c}(\beta, \omega_l) \qquad s.t. \quad p_{\mathrm{R}}^{\omega_l} UB(\beta, \lambda_m^{\omega_l}) \leq \nu, \tag{2.32}$$

for $m \in \mathbb{Z}^+$. Then there exists a $\beta^\star \geq 0$ such that

$$\lim_{m\to\infty} \beta_m^G = \lim_{m\to\infty} \beta_m^F = \beta^\star,$$

where the β_m^F are the optimal staffing factors for problem (2.26) for all m.

Proof. We again start with the approximate problem (2.27), which can be transformed identically into

$$\max_{d\in\mathcal{D}} \left\{ \left(\tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_l}) - \frac{\lambda^{\omega_l} - \lambda^{\omega_1}}{\lambda^{\omega_{l+1}} - \lambda^{\omega_1}} \right) p^{\omega_l}(d) + \frac{r - \lambda^{\omega_1}}{\lambda^{\omega_{l+1}} - \lambda^{\omega_1}} \right\} \tag{2.33}$$

by Lemma 2.4.1, with some variable substitution. Lemma 2.4.3 proves that for any $\delta > 0$, there exists an $\bar{m}$ such that for all $m > \bar{m}$, (2.33) is greater than

$$\max_{d\in\mathcal{D}} \left\{ \sum_{k=1}^{n} p^{\omega_k}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \right\} - \delta.$$

Choosing $\delta = \epsilon - \frac{r-\lambda^{\omega_1}}{\lambda^{\omega_{l+1}} - \lambda^{\omega_1}}$, we have

$$\max_{d\in\mathcal{D}} \left\{ \left(\tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_l}) - \frac{\lambda^{\omega_l} - \lambda^{\omega_1}}{\lambda^{\omega_{l+1}} - \lambda^{\omega_1}} \right) p^{\omega_l}(d) \right\}$$
$$> \max_{d\in\mathcal{D}} \left\{ \sum_{k=1}^{n} p^{\omega_k}(d) \tilde{\alpha}(\beta, \lambda_m^{\omega_l}, \lambda_m^{\omega_k}) \right\} - \epsilon, \tag{2.34}$$

for any $\beta \geq 0$. If $\beta = \beta_m^F$, the right-hand side of (2.34) has to be 0. Otherwise, the constraint in problem (2.26) is not binding and β_m^F cannot be optimal.

Therefore, we have

$$\max_{d\in\mathcal{D}}\left\{\left(\tilde{\alpha}(\beta_m^F,\lambda_m^{\omega_l},\lambda_m^{\omega_l})-\frac{\lambda^{\omega_l}-\lambda^{\omega_1}}{\lambda^{\omega_{l+1}}-\lambda^{\omega_1}}\right)p^{\omega_l}(d)\right\}>0,$$

and thus

$$\tilde{\alpha}(\beta_m^F,\lambda_m^{\omega_l},\lambda_m^{\omega_l})-\frac{\lambda^{\omega_l}-\lambda^{\omega_1}}{\lambda^{\omega_{l+1}}-\lambda^{\omega_1}}>0.$$

Let $\mathbf{p}_{\text{R}}$ be a maximizer of (2.33) when $\beta=\beta_m^F$. With a positive coefficient, $p_{\text{R}}^{\omega_l}$ should take the largest feasible $p^{\omega_l}(d)$ as shown in the third column of Table 2.1. The values of $p_{\text{R}}^{\omega_1}$ and $p_{\text{R}}^{\omega_{l+1}}$ are then computed, and $\mathbf{p}_{\text{R}}$ is determined with the probabilities in all other scenarios being 0. Recall that the distribution given by $\mathbf{p}_{\text{R}}$ is also the unique optimal solution to problem (2.27).

As a linear programming (LP) problem, the left-hand side of the constraint in model (2.26) differs from problem (2.27) only in the objective coefficients, and the deviation goes to zero as m goes to infinity. We know that if there exists a unique optimal solution to an LP model, it will remain optimal after some changes in the objective coefficients as long as they are small enough. Hence, we can rewrite problem (2.26) equivalently as

$$\min_{\beta\geq 0}\bar{c}(\beta,\omega_l)\qquad\text{s.t.}\quad\sum_{k=1}^{n}p_{\text{R}}^{\omega_k}\tilde{\alpha}(\beta,\lambda_m^{\omega_l},\lambda_m^{\omega_k})\leq\epsilon$$

when m is sufficiently large. Defining $\nu=\epsilon-\sum_{k=l+1}^{n}p_{\text{R}}^{\omega_k}$, the result then follows directly from Theorem 2.2.4. □

Example 4. Consider problem (2.22) and suppose the arrival rate Λ has state space $\{100,200,400,700\}$ with $r=250$. Set the QoS requirement to $\epsilon=0.30$.

We start with identification of the key scenario. From the first two columns of Table 2.1, we know $\lambda^{\omega_l} = 400$ because $\frac{250-100}{700-100} < \epsilon < \frac{250-100}{400-100}$. Hence $p_{\mathrm{R}}^{\omega_l} = 0.50$ and $\nu = 0.30$ according to the last two cells of the middle row. Next, we consider problem (2.32) and repeat procedures used in previous examples. We solve for β in $0.50 \cdot UB(\beta, 400) = 0.30$, and plug $\beta = 0.387$ into the square-root safety staffing rule, which gives the solution $s = 408$ with $400 + 0.387 \cdot \sqrt{400} = 407.74$. □

2.5 Computational Results

Now that we have approximately solved both Level II and Level III problems by virtue of asymptotic optimality results, we want to compare these models and evaluate how much benefit might be obtained by knowing the entire arrival-rate distribution. In addition, we are interested in exploring the effect of changing the assumptions of the UMD model.

2.5.1 Value of Information

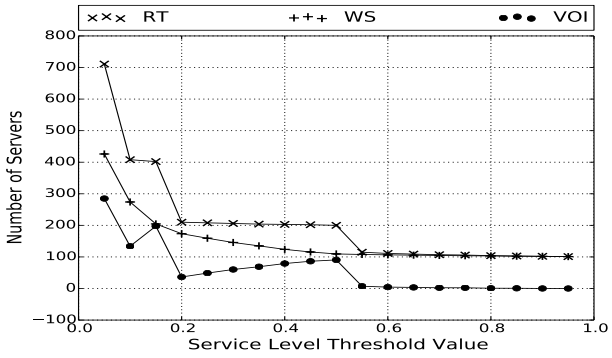
One possible consequence of formulating a UMD model is the associated optimal solution turns out to be infeasible for the Level II problem given by the true arrival-rate distribution. So, there is no consistent way to quantify the VOI gained by applying the Level II model instead without introducing some penalty on QoS violation. Due to the vagaries of assigning such penalties we do not explore this type of VOI herein. Rather we compute VOI for the robust formulation since its calculation does not require any model modification.

We define VOI for the robust model to be the difference between its optimal value and the expectation of the wait-and-see Level II solution over some meta-distribution D. This VOI is formally given by

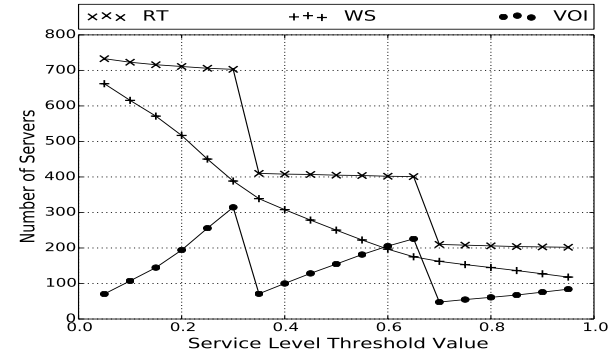
$$\text{VOI} := \min_{s \in \bigcap_{d \in \mathcal{D}} X[\mathbf{p}(d)]} c(s) - \mathbb{E}_D \left[\min_{s \in X[\mathbf{p}(D)]} c(s) \right],$$

where $X(\cdot)$ denotes the feasible region of a Level II problem for a certain arrival-rate distribution.

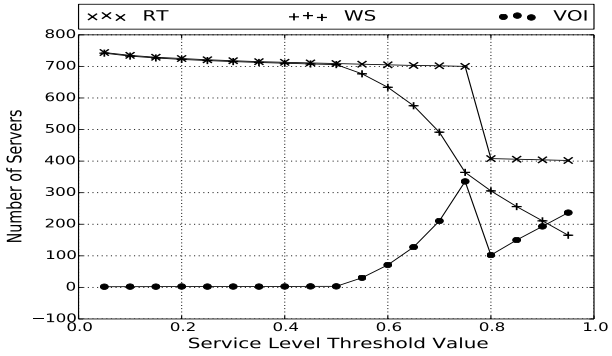
Again, we assume nature is apathetic and D is uniformly distributed. We are then able to estimate the expected wait-and-see solution by averaging optimal values of a sufficiently large number of instances (2.6), each of which is parameterized with a distribution randomly selected from $\mathcal{D}$. There are different ways of sampling points uniformly from a polytope, and here we use the hit-and-run algorithm (see, for example, Montiel and Bickel [46]). In the discussion below, we consider the approximate robust solution, denoted by RT, derived in the same manner as in Example 4. We also consider the mean wait-and-see solution which is an estimate of the mean number of servers for a collection of Level II models, where every sample instance is solved with the procedure in Example 2. In Figure 2.5 and Figure 2.6 we display the results of the VOI calculation with various parameter settings and for $c(s) = s$. In general, we keep the gap between λ^{ω_k} and $\lambda^{\omega_{k+1}}$ greater than $\sqrt{16 \cdot \lambda^{\omega_k}}$ for $\lambda^{\omega_1} \geq 100$ so that the delay probability can be approximated by 0 or 1, except in the key scenario. The standard errors are all less than 0.41 with a sample size of 1.6×10^5, so the error bars are imperceptible in the charts.



(a) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 150$

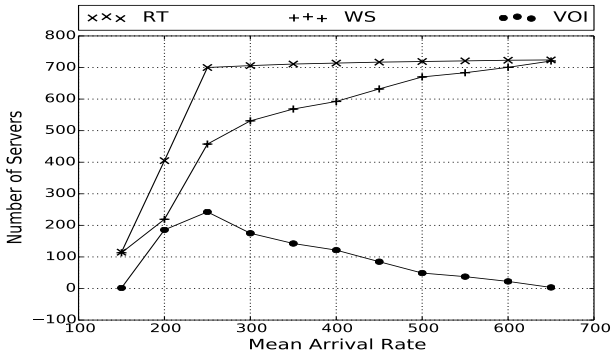


(b) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 300$

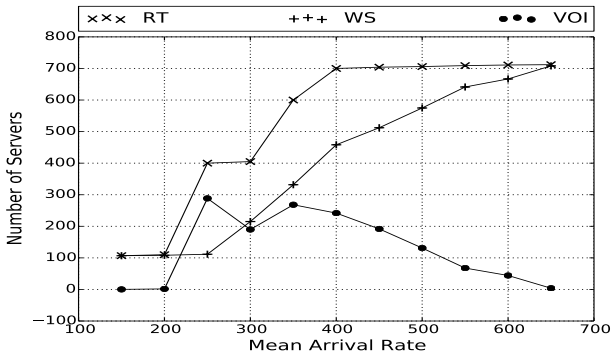


(c) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 550$

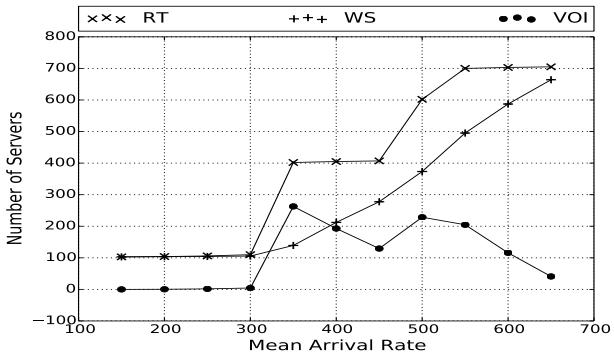
Figure 2.5: The charts depict VOI fluctuation with different values of r, where the approximate robust solution (RT) is compared with the corresponding mean wait-and-see solution (WS).



(a) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.25$



(b) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.50$



(c) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.75$

Figure 2.6: The charts depict VOI fluctuation with different values of ϵ, where the approximate robust solution (RT) is compared with the corresponding mean wait-and-see solution (WS).

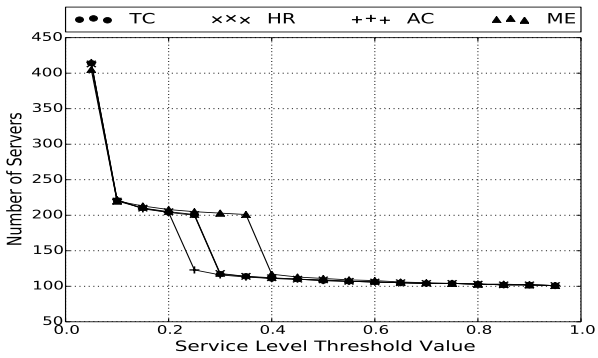
The zig-zag pattern of the VOI in Figure 2.5 and Figure 2.6 is related to the dramatic decline in the value of the approximate robust solution, which occurs whenever the worst case key scenario drops in value. On the other hand, it is not surprising that both the robust and the wait-and-see curves appear to be declining consistently with an increase in the service level threshold value or a decrease in the mean arrival rate. We also recognize two cases where the VOI is small. If the mean arrival rate is close to the lowest possible arrival rate, the VOI is not going to be significant for a problem with a large ϵ since the QoS constraint can always be easily satisfied by a low staffing level. On the other hand, an extremely large value of r can lead to very limited VOI if the service level threshold value is small. This is because the QoS standard is so high that any arrival-rate distribution with the given r will require a large number of agents.

2.5.2 Extensions

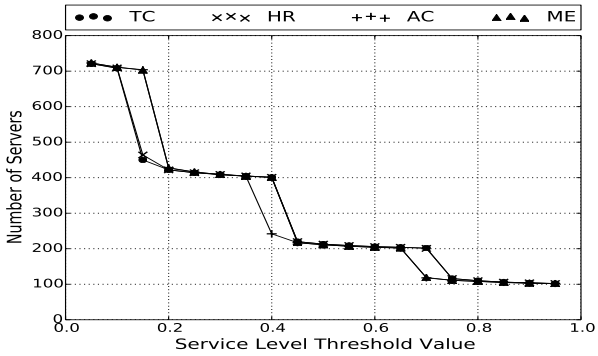
We consider the sample points generated by the hit-and-run algorithm. Notice that their arithmetic mean position yields an estimate of the centroid, which suggests a way of approximating $\mathbf{p}_{\mathrm{U}}$ for the UMD model in the case of $n > 4$. In Figure 2.7 and Figure 2.8, we let TC (for true centroid) denote the approximately optimal staffing level for a UMD model with $c(s) = s$, which is solved by the procedure from Example 3. In other words, this is the approximately optimal value of the Level II problem defined by the true centroid of $\mathcal{D}$. If we replace the true centroid with the hit-and-run estimate of the centroid,

the approximately optimal value of the Level II problem parameterized with the new arrival-rate distribution may deviate from the true centroid solution. As a matter of fact, the discrepancy is sometimes non-negligible as seen in Figure 2.7(c) and Figure 2.8(c).

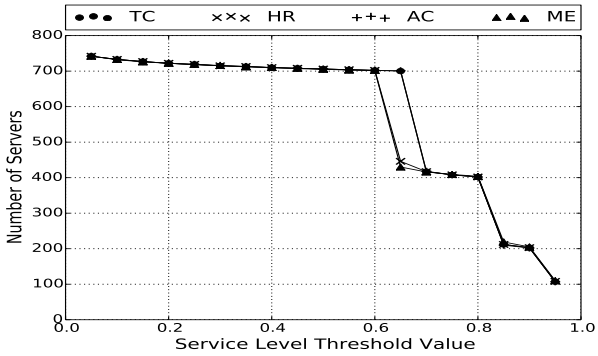
We may also extend our computational experiments to other special points inside the polytope $\mathcal{D}$. For example, when $c(s) = s$ we can also compute the approximate solutions to Level II staffing problems where the arrival-rate distribution is either the analytic center of $\mathcal{D}$ or the maximum entropy distribution in $\mathcal{D}$. We observe from Figure 2.7 and Figure 2.8 that the four series of results are often similar, especially when a relatively small r combines with a large ϵ, or ϵ is small while r is large compared to all possible arrival rates. Hence, the analytic center and maximum entropy solutions can provide reasonable estimates of the UMD solution as well in some cases, and the computation of the two former points requires much less effort than the hit-and-run procedure.



(a) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 150$

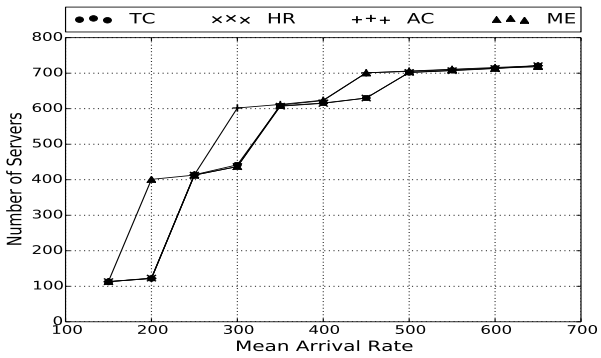


(b) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 300$

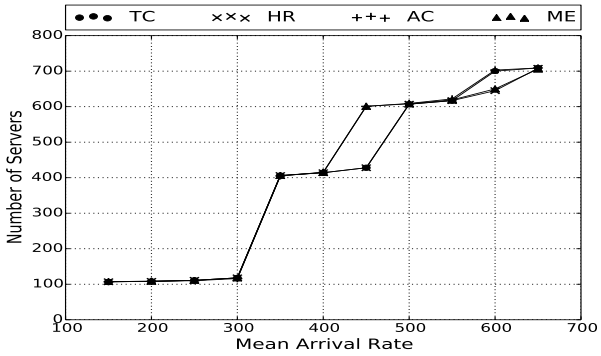


(c) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 200, 400, 700)$, $r = 550$

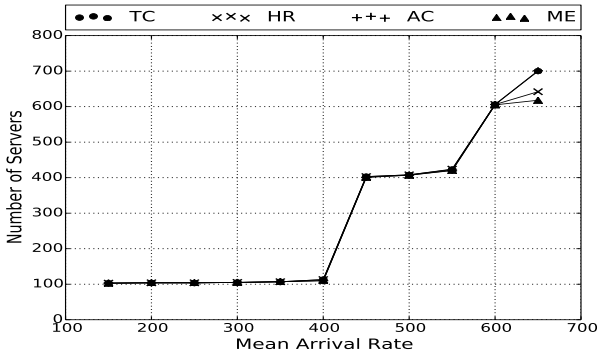
Figure 2.7: The charts depict approximate optimal staffing levels with different values of r for the UMD model (TC) and its extensions, in which we replace the true centroid with the hit-and-run estimate (HR), the analytic center (AC) and the maximum entropy distribution (ME) respectively.



(a) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.25$



(b) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.50$



(c) $(\lambda^{\omega_1}, \lambda^{\omega_2}, \lambda^{\omega_3}, \lambda^{\omega_4}) = (100, 400, 600, 700)$, $\epsilon = 0.75$

Figure 2.8: The charts depict approximate optimal staffing levels with different values of ϵ for the UMD model (TC) and its extensions, in which we replace the true centroid with the hit-and-run estimate (HR), the analytic center (AC) and the maximum entropy distribution (ME) respectively.

Chapter 3

Staffing Multi-Station Service Systems with Joint QoS Constraints

3.1 Introduction

In this chapter, we focus on systems where customers are grouped into classes based on their service requests, which can be accomplished only by dedicated stations. We assume there are L ($L \geq 2$) customer classes in total, where the queue length process for each class corresponds to an $M/M/s_j$ queue with $j = 1, \ldots, L$. The system structure is shown in Figure 3.1.

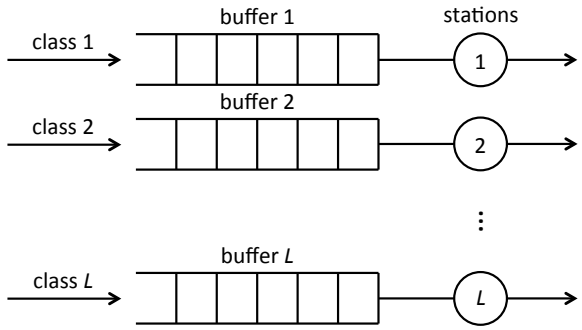


Figure 3.1: Structure of a Multi-Station Service System

We consider the staffing problem proposed by Zan et al. [60] for multi-station systems, where the single-station QoS constraint in Chapter 2 is extended to a joint one, under which we minimize the total staffing cost induced.

Compared to enforcing an individual QoS constraint for each customer class, implementing the joint constraint provides a higher-level control in operations by allowing smart allocation of the service delay probability across all stations. The arrival rates are known constants in Section 3.2, while they are discrete random variables following given distributions in Section 3.3. We are specifically interested in systems that are large in scale in the latter case.

3.2 Systems with Deterministic Arrival Rates

We assume the service rate for each server is 1 without loss of generality. Recall the service delay probability for an $M/M/s$ queue with an arrival rate of λ, denoted by $\alpha(s, \lambda)$, is computed by the Erlang-C formula for $s > \lambda$. We again use its continuous extension $\bar{\alpha}(s, \lambda)$ in the following analysis as in Chapter 2. Let λ_j be the arrival rate and c_j ($c_j > 0$) be the unit staffing cost for station j. Assuming independent queueing dynamics, we formulate the multi-station staffing problem as

$$\min_{\mathbf{s}} \quad \sum_{j=1,\dots,L} c_j s_j \tag{3.1a}$$

$$\text{s.t.} \quad \prod_{j=1,\dots,L} (1 - \bar{\alpha}(s_j, \lambda_j)) \geq 1 - \epsilon \tag{3.1b}$$

$$s_j > \lambda_j, \quad \forall j \in \{1, \dots, L\}, \tag{3.1c}$$

where $\epsilon \in (0, 1)$ represents the joint QoS threshold value, and $\mathbf{s} := (s_1, \dots, s_L)$. We can interpret constraint (3.1b) as the stationary probability that all queues are empty should be no less than $1 - \epsilon$. The requirement is strong in the QoS sense, but it may be applied to systems demanding break time.

Proposition 3.2.1. *There exists a unique optimal solution to problem* (3.1).

Proof. We know $\bar{\alpha}(s_j, \lambda_j)$ is continuous and strictly decreasing in s_j with an infimum of 0, and it goes to 1 when s_j approaches λ_j from above. Therefore, being a minimization problem, (3.1) has optimal solutions that are achieved when (3.1b) is binding.

Since there are no equality constraints and only one active inequality constraint for any potential optimal solution, the problem satisfies the linear independence constraint qualification, and thus the Karush-Kuhn-Tucker (KKT) conditions are necessary for optimality. Let τ be the dual variable for (3.1b) and $\nu_1, \ldots, \nu_L$ for (3.1c) in optimality. According to complementary slackness, $\nu_j = 0$ for $j = 1, \ldots, L$. Hence, the stationarity condition is

$$-c_j = \tau \cdot \frac{\partial \bar{\alpha}(s_j, \lambda_j)}{\partial s_j} \cdot \prod_{j' \in \{1,\ldots,L\} \setminus \{j\}} (1 - \bar{\alpha}(s_{j'}, \lambda_{j'})) \tag{3.2}$$

for all j. Suppose there are two different optimal solutions $\mathbf{s}^\star$ and $\mathbf{s}'$ to problem (3.1). Then there must exist $j_1 \in \{1, \ldots, L\}$, $j_2 \in \{1, \ldots, L\}$ and $j_1 \neq j_2$ such that $s^\star_{j_1} > s'_{j_1}$ and $s^\star_{j_2} < s'_{j_2}$. Writing out equation (3.2) for $j = j_1$ and $j = j_2$ respectively and taking the ratio, we obtain

$$\frac{c_{j_1}}{c_{j_2}} = \frac{{}^{\partial \bar{\alpha}\left(s_{j_1}, \lambda_{j_1}\right)}\!/_{\partial s_{j_1}}}{{}^{\partial \bar{\alpha}\left(s_{j_2}, \lambda_{j_2}\right)}\!/_{\partial s_{j_2}}} \cdot \frac{1 - \bar{\alpha}\left(s_{j_2}, \lambda_{j_2}\right)}{1 - \bar{\alpha}\left(s_{j_1}, \lambda_{j_1}\right)}. \tag{3.3}$$

Recalling the strict monotonicity of $\bar{\alpha}(s_j, \lambda_j)$ in s_j, we know

$$0 < \frac{1 - \bar{\alpha}\left(s^\star_{j_2}, \lambda_{j_2}\right)}{1 - \bar{\alpha}\left(s^\star_{j_1}, \lambda_{j_1}\right)} < \frac{1 - \bar{\alpha}\left(s'_{j_2}, \lambda_{j_2}\right)}{1 - \bar{\alpha}\left(s'_{j_1}, \lambda_{j_1}\right)}. \tag{3.4}$$

In addition, the convexity of $\bar{\alpha}(s_j, \lambda_j)$ in s (see [35]) leads to

$$0 < \left. \frac{\partial\bar{\alpha}\left(s_{j_1},\lambda_{j_1}\right)/\partial s_{j_1}}{\partial\bar{\alpha}\left(s_{j_2},\lambda_{j_2}\right)/\partial s_{j_2}} \right|_{\mathbf{s}=\mathbf{s}^\star} \leq \left. \frac{\partial\bar{\alpha}\left(s_{j_1},\lambda_{j_1}\right)/\partial s_{j_1}}{\partial\bar{\alpha}\left(s_{j_2},\lambda_{j_2}\right)/\partial s_{j_2}} \right|_{\mathbf{s}=\mathbf{s}'}. \tag{3.5}$$

Given (3.4) and (3.5), we conclude (3.3) cannot hold for both $\mathbf{s}^\star$ and $\mathbf{s}'$, and the assumption of multiple optimal solutions is therefore false. □

3.3 Systems with Uncertain Arrival Rates

Now we assume random arrival rates for all stations, which are possibly correlated with each other. The service rates are still assumed to be 1. Let Λ_j be the arrival rate for the customer class j, and we define its state space to be $\left\{\lambda_{(j,1)}, \ldots, \lambda_{(j,n_j)}\right\}$. Taking the asymptotic view as in Chapter 2, we consider a sequence of systems for which the arrival rate is $\Lambda_j^m \in \left\{\lambda_{(j,1)}^m, \ldots, \lambda_{(j,n_j)}^m\right\}$ for $m \in \mathbb{Z}^+$, where $\lambda_{(j,k_j)}^m = m\lambda_{(j,k_j)}$ for $k_j \in \{1, \ldots, n_j\}$.

Following the analysis in the single-station case, we want to determine the key scenario, apply the square-root safety staffing rule and replace $\bar{\alpha}$ with the JVLZ upper bound. The idea of searching for a joint key scenario relies again on the fact that given square-root safety factors, there can be non-trivial service delay probabilities in only one case under the specified scaling. That is, the probabilities converge to either 0 or 1 with $m \to \infty$ depending on whether arrival-rate realizations are lower or higher than the key arrival rates. However, unlike in a single-station problem, there seems to be no obvious direction for searching, and there can exist multiple scenarios that are of interest. The worst case is to find such scenarios by enumeration, which should not be hard

given dimensions of real-life problems. A detailed example of implementing the procedure is presented in [60].

Unable to prove the convexity of the JVLZ bound, we skip the step of parameterizing the problem with the square-root safety staffing rule. Instead, we consider the model

$$\min_{\mathbf{s}} \quad \sum_{j=1,\dots,L} c_j s_j \tag{3.6a}$$

$$\text{s.t.} \quad \sum_{E\subseteq\{1,\dots,L\}} \left[p(E) \prod_{j'\in E} \left(1 - \bar{\alpha}\left(s_{j'}, \lambda^m_{\left(j',k^\star_{j'}\right)} \right) \right) \right] \geq 1-\epsilon \tag{3.6b}$$

$$s_j > \lambda^m_{\left(j,k^\star_j\right)}, \quad \forall j \in \{1,\dots,L\}, \tag{3.6c}$$

where $\lambda^m_{(j,k^\star_j)}$ is the class j arrival rate in the predetermined key scenario, and $p(E)$ is the total probability of observing scenarios that satisfy $\Lambda^m_j = \lambda^m_{\left(j,k^\star_j\right)}$ if $j \in E$ and $\Lambda^m_j < \lambda^m_{\left(j,k^\star_j\right)}$ otherwise.

Proposition 3.3.1. *When $L = 2$, there exists an $M \in \mathbb{Z}^+$ such that the optimal solution to problem* (3.6) *is unique for $m \geq M$.*

Proof. We conclude optimal solutions exist, and they must satisfy the KKT conditions using the same reasoning as in the proof of Proposition 3.2.1. The

stationarity condition requires

$$\frac{c_1}{c_2} = \frac{\partial \bar{\alpha}\left(s_1, \lambda^m_{\left(1,k_1^\star\right)}\right)}{\partial s_1} \cdot \left[\frac{\partial \bar{\alpha}\left(s_2, \lambda^m_{\left(2,k_2^\star\right)}\right)}{\partial s_2}\right]^{-1} \cdot \frac{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_2, \lambda^m_{\left(2,k_2^\star\right)}\right)\right) + p(\{1\})}{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_1, \lambda^m_{\left(1,k_1^\star\right)}\right)\right) + p(\{2\})}. \quad (3.7)$$

Suppose both $\mathbf{s} = \mathbf{s}^\star$ and $\mathbf{s} = \mathbf{s}'$ are optimal for (3.6), and $\mathbf{s}^\star \neq \mathbf{s}'$. Without loss of generality, we assume $s_1^\star > s_1'$ and $s_2^\star < s_2'$, and therefore

$$\frac{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_2^\star, \lambda^m_{\left(2,k_2^\star\right)}\right)\right) + p(\{1\})}{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_1^\star, \lambda^m_{\left(1,k_1^\star\right)}\right)\right) + p(\{2\})} < \frac{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_2', \lambda^m_{\left(2,k_2^\star\right)}\right)\right) + p(\{1\})}{p(\{1,2\}) \cdot \left(1 - \bar{\alpha}\left(s_1', \lambda^m_{\left(1,k_1^\star\right)}\right)\right) + p(\{2\})}$$

as long as $p(\{1,2\}) \neq 0$. Notice that $p(\{1,2\}) = 0$ means the key scenario occurs with a probability of 0, which contradicts its optimality for a sufficiently large m. The convexity of $\bar{\alpha}\left(s_j, \lambda^m_{\left(j,k_j^\star\right)}\right)$ in s_j then suggests (3.7) is violated by either $\mathbf{s} = \mathbf{s}^\star$ or $\mathbf{s} = \mathbf{s}'$. □

Proposition 3.3.1 cannot be directly generalized to problems with $L > 2$, and it requires further exploration to determine whether there are multiple optimal solutions in such cases.

3.4 Future Research

If we know there is a unique solution satisfying the KKT conditions, and it is indeed optimal, we can numerically find it by nonlinear programming algorithms. However, we may encounter the problem of numerical instability, so we wish to substitute the Erlang-C formula with the JVLZ upper bound. If the JVLZ upper bound is proved to be convex in the safety factor, the current uniqueness results still hold, and we can then prove or disprove asymptotic optimality of the approximate solution. It is also desirable to consider an additional level of stochasticity and analyze Level III staffing problems in the multi-station scenario.

Chapter 4

Strategic Pricing of Service Systems with Uncertain Arrival Rates

4.1 Introduction

We consider the balking model for a first-come-first-served $M/M/1$ system, where arriving customers can leave before entering the queue. Customers are assumed to be risk neutral, and they make decisions so as to maximize their expected individual utility. The problem was studied first by Naor [47] for observable queues in which queue length information is provided to customers upon arrival, and there has been substantial literature on modeling such game-theoretic problems since his seminal work. In particular, Edelson and Hildebrand [25] extend the joining or balking analysis to the context of unobservable queues that hide queue length from all customers. In this chapter, we explore how classic results from the aforementioned work are affected by introducing stochastic arrival rates.

The arrival rate is now assumed to be a non-degenerate positive random variable Λ, while the service rate is a given constant μ. We assume nature picks a realization of Λ before the start of the analysis horizon, and customers then arrive at the system following a Poisson process defined by the chosen rate, say,

λ. However, customers and other decision makers are ignorant about the value of λ the entire time, although they have complete knowledge of the distribution of Λ. The setting is realistic in the sense that potential service demands may vary with unpredictable factors and are therefore difficult to estimate. Except for this, we adopt Naor's assumptions for observable systems and Edelson and Hildebrand's assumptions for unobservable systems. We carry over the notation such that R is the monetary value of benefits obtained by a customer after service completion, and C is the cost per unit of time per customer for waiting in the queue. Our goal is to investigate how the optimal joining strategies shift with different standpoints of individual customers, the social optimizer and the profit maximizer, and furthermore, what pricing schemes can induce customers to follow these strategies.

A detailed review on game-theoretic models with customer queueing behavior is presented by Hassin and Haviv [33], while all the single-server joining or balking problems included do not involve uncertainty of potential arrival rates. For example, Chen and Frank [20] allow adjusting the entering fee of a queue based on real-time queue length, and they find the price that maximizes the system profit does not maximize social welfare when customers have heterogeneous service valuations. Besbes and Maglaras [13] study dynamic pricing in the cases where underlying dynamics of non-homogeneous Poisson arrivals cannot be formulated by a precise model. Using fluid approximations, they obtain near-optimal policies for large-scale systems with a slow-varying market size. Afèche and Ata [1] develop Bayesian pricing policies for systems

with unknown demand scenarios that are defined by the percentage of impatient customers. Haviv and Randhawa [34] analyze demand-independent static pricing for unobservable systems, which performs considerably well in revenue optimization for some distributions of customer valuations. Compared with their problem setting, we also apply static pricing to regulating the system, while we assume certain knowledge of the demand, and we focus on generalizing Naor's analysis to include arrival-rate uncertainty for both observable (Section 4.2) and unobservable (Section 4.3) queues. Last but not least, instead of uncertain arrival rates, Zheng [61] considers random service rates in unobservable systems when analyzing balking behavior of optimistic and pessimistic customers. To the best of our knowledge, there has been no work on queueing systems with strategic customer behavior and a known arrival-rate distribution.

4.2 Observable Queues

Our analysis in this section adheres to Naor's framework for observable queues. For any arriving customer, the optimal threshold of joining is

$$\tilde{n}_e = \left\lfloor \frac{R\mu}{C} \right\rfloor,$$

which is exactly the same as when Λ is a constant since customers do not require knowledge of the arrival rate when the queue is observable.

Let $\tilde{n}_s \in \mathbb{Z}^+$ denote the threshold of joining that maximizes the total benefit received by all customers. Recall they are homogeneous and have

exponential inter-arrival and service times. Mimicking Hassin's analysis for problems with deterministic arrival rates in [32], we find $\tilde{n}_s$ coincides with the optimal reneging threshold from the perspective of individual customers under a preemptive last-come first-served queueing discipline. Therefore, $\tilde{n}_s$ is the largest n that satisfies

$$R \cdot \mathbb{E}_\Lambda \left[\frac{1-\rho}{1-\rho^{n+1}} \right] - \frac{C}{\mu} \cdot \mathbb{E}_\Lambda \left[\frac{n}{1-\rho} - \frac{(n+1)\rho(1-\rho^n)}{(1-\rho)(1-\rho^{n+1})} \right] \geq 0, \tag{4.1}$$

where $\rho := \Lambda/\mu$ and $\rho \neq 1$ (see equation (2.5) in [33]). We do not separately discuss the case of $\rho = 1$ because the results in observable queues remain true when taking the limit as $\rho \to 1$, as pointed out in [33].

Proposition 4.2.1. *Define* $f(v, \Lambda) = \frac{v}{1-\rho} - \frac{(v+1)\rho(1-\rho^v)}{(1-\rho)(1-\rho^{v+1})}$ *and* $g(v, \Lambda) = \frac{1-\rho}{1-\rho^{v+1}}$ *for* $v \in \mathbb{R}^+$. *Let* $v = v_s$ *be such that*

$$\frac{\mathbb{E}_\Lambda \left[f(v_s, \Lambda) \right]}{\mathbb{E}_\Lambda \left[g(v_s, \Lambda) \right]} = \frac{R\mu}{C}.$$

Then $\tilde{n}_s = \lfloor v_s \rfloor$.

Proof. With $g(v, \lambda) > 0$ for any constant $\lambda > 0$, we know $\mathbb{E}_\Lambda[g(v, \Lambda)] > 0$, and thus (4.1) can be written equivalently as

$$\frac{R\mu}{C} \geq \frac{\mathbb{E}_\Lambda \left[f(v, \Lambda) \right]}{\mathbb{E}_\Lambda \left[g(v, \Lambda) \right]}. \tag{4.2}$$

Let $\rho_\lambda := \lambda/\mu$. We know

$$\frac{\partial f(v, \lambda)}{\partial v} = \frac{\rho_\lambda^{v+1} \log \rho_\lambda (1+v) - \rho_\lambda^{v+1} + 1}{(1-\rho_\lambda^{v+1})^2} \geq 0,$$

so $f(v,\lambda)$ is non-decreasing in v. We also know $g(v,\lambda)$ is decreasing in v. Besides, monotonicity is preserved under expectation. Hence, the right-hand side of (4.2) is strictly increasing in v, and $\lfloor v_s \rfloor$ is the largest n that satisfies (4.1). □

Theorem 4.2.2. $\tilde{n}_s \leq \tilde{n}_e$.

Proof. We first compute the difference between $\mathbb{E}_\Lambda[f(v,\Lambda)]$ and $v\mathbb{E}_\Lambda[g(v,\Lambda)]$:

$$\begin{aligned}\mathbb{E}_\Lambda[f(v,\Lambda)] - v\mathbb{E}_\Lambda[g(v,\Lambda)] &= \mathbb{E}_\Lambda\left[\frac{v}{1-\rho} - \frac{(v+1)\rho(1-\rho^v)}{(1-\rho)(1-\rho^{v+1})} - v\cdot\frac{1-\rho}{1-\rho^{v+1}}\right] \\ &= \mathbb{E}_\Lambda\left[\frac{\rho^{v+1} - v\rho^2 + v\rho - \rho}{(1-\rho)(1-\rho^{v+1})}\right].\end{aligned}$$

For any constant $\lambda > 0$, we define $\rho_\lambda = \lambda/\mu$ and $h(v,\lambda) = \rho_\lambda^{v+1} - v\rho_\lambda^2 + v\rho_\lambda - \rho_\lambda$. Taking the derivative of $h(v,\lambda)$ with respect to v, we have

$$\frac{\partial h(v,\lambda)}{\partial v} = \rho_\lambda^{v+1}\log\rho_\lambda - \rho_\lambda^2 + \rho_\lambda \geq \rho_\lambda^v(\rho_\lambda - 1) - \rho_\lambda(\rho_\lambda - 1) \geq 0$$

for all ρ_λ when $v \geq 1$. With $h(1,\lambda) = 0$, we conclude $h(v,\lambda) \geq 0$ and $\mathbb{E}_\Lambda[f(v,\Lambda)] \geq v\mathbb{E}_\Lambda[g(v,\Lambda)]$ for $v \geq 1$ because expectation preserves the inequality. The right-hand side of (4.2) is thus no less than v for $v \geq 1$. Since $v = v_s$ satisfies (4.2), $\frac{R\mu}{C} \geq v_s$, and $\tilde{n}_e \geq \tilde{n}_s$. □

To induce customers to follow the socially optimal joining strategy, we allow the system to charge a static entering fee to each customer that joins the system. The collected fee is regarded as a transfer payment from the view of social welfare. We now consider a profit maximizing firm. If the firm can

set the joining threshold n, then the profit rate is given by (see equation (2.9) in [33]):

$$Z_1(n) := \mathbb{E}_\Lambda\left[\Lambda\frac{1-\rho^n}{1-\rho^{n+1}}\right]\left(R-\frac{nC}{\mu}\right).$$

We want to compute $\tilde{n}_m \in \arg\max_{n\in\mathbb{Z}^+}\{Z_1(n)\}$ so as to determine the entering fee amount that induces the maximum total profit.

Proposition 4.2.3. *Define* $u(v,\Lambda) = \frac{1-\rho^{v-1}}{1-\rho^v}$ *and* $w(v,\Lambda) = \frac{(1-\rho)^2\rho^{v-1}}{(1-\rho^{v+1})(1-\rho^v)}$ *for* $v \in \mathbb{R}^+$. *Let* $v = v_m$ *be such that*

$$v_m + \frac{\mathbb{E}_\Lambda\left[u(v_m,\Lambda)\right]}{\mathbb{E}_\Lambda\left[w(v_m,\Lambda)\right]} = \frac{R\mu}{C}. \tag{4.3}$$

Then $\tilde{n}_m = \lfloor v_m \rfloor$.

Proof. By definition of $\tilde{n}_m$, we have $Z_1(\tilde{n}_m) \geq Z_1(\tilde{n}_m - 1)$ and $Z_1(\tilde{n}_m) \geq Z_1(\tilde{n}_m + 1)$, which lead to

$$\tilde{n}_m + \frac{\mathbb{E}_\Lambda\left[u(\tilde{n}_m,\Lambda)\right]}{\mathbb{E}_\Lambda\left[w(\tilde{n}_m,\Lambda)\right]} \leq \frac{R\mu}{C} \leq \tilde{n}_m + 1 + \frac{\mathbb{E}_\Lambda\left[u(\tilde{n}_m+1,\Lambda)\right]}{\mathbb{E}_\Lambda\left[w(\tilde{n}_m+1,\Lambda)\right]}. \tag{4.4}$$

Let $\rho_\lambda := \lambda/\mu$. We know

$$\frac{\partial u(v,\lambda)}{\partial v} = \frac{\log\rho_\lambda(\rho_\lambda^v - \rho_\lambda^{v-1})}{(1-\rho_\lambda^v)^2} \geq 0,$$

and

$$\frac{\partial w(v,\lambda)}{\partial v} = \frac{\rho_\lambda^{v-1}\log\rho_\lambda(1-\rho_\lambda)^2(1-\rho_\lambda^{2v+1})}{(1-\rho_\lambda^{v+1})^2(1-\rho_\lambda^v)^2} \leq 0$$

for $v \in \mathbb{R}^+$. Hence, the term $v + \frac{\mathbb{E}_\Lambda[u(v,\Lambda)]}{\mathbb{E}_\Lambda[w(v,\Lambda)]}$ is non-decreasing in v, and it ranges from 1 to ∞ for $v \geq 1$. Since we assume $R\mu \geq C$, we can always solve for v_m in (4.3) and then round it down to find $\tilde{n}_m$. □

Corollary 4.2.4. $\tilde{n}_m \leq \tilde{n}_e$.

Proof. The corollary follows from equation (4.3) and the fact that $\mathbb{E}_\Lambda [u(v_m, \Lambda)]$ and $\mathbb{E}_\Lambda [w(v_m, \Lambda)]$ are positive. □

It is desirable to discover the relationship between $\tilde{n}_s$ and $\tilde{n}_m$, and numerical evidence suggests $\tilde{n}_m \leq \tilde{n}_s$, but for some parameters numerical error makes it difficult to validate this claim. This result is an open conjecture right now.

4.3 Unobservable Queues

Unlike in the previous case, customers may not observe the queue length in unobservable systems, and they are concerned about the arrival rate when they decide whether to balk. Let $W(\cdot)$ denote the expected waiting time for a given arrival rate. For an arriving customer, the equilibrium strategy is to join the queue with probability of $\tilde{q}_e$, where $\tilde{q}_e := \min\{\bar{q}_e, 1\}$, and $\bar{q}_e$ is such that $C\mathbb{E}_\Lambda[W(\bar{q}_e\Lambda)] = R$, or

$$\mathbb{E}_\Lambda \left[\frac{1}{\mu - \bar{q}_e\Lambda}\right] = \frac{R}{C}. \tag{4.5}$$

We then consider charging each joining customer an entering fee p. For individual decision making, this is equivalent with reducing the service benefit from R to $R - p$, which results in a shift of the percentage of joining from $\tilde{q}_e$ in equilibrium. Hence, no customers are willing to join if $p \geq R - C\mathbb{E}_\Lambda[W(0)]$, that is, $p \geq R - \frac{C}{\mu}$. We use $q(p)$ to denote the deviated joining probability

associated with p. Again, we view p as a transfer payment when computing social welfare, and the social optimization model is

$$\max_{q(p)\in[0,1]\cap\left[0,\mu/\bar{\lambda}\right)} \mathbb{E}_{\Lambda}[q(p)\Lambda(R - CW(q(p)\Lambda))] \tag{4.6}$$

with $\bar{\lambda}$ as the maximum possible arrival rate. We define $Z_2(\bar{q}) = \mathbb{E}_{\Lambda}[\bar{q}\Lambda(R - CW(\bar{q}\Lambda))]$ for $\bar{q} \in \left[0, \mu/\bar{\lambda}\right)$, which is the total social benefit with the joining percentage $\bar{q}$ if $\bar{q} \leq 1$.

Proposition 4.3.1. *Let $\bar{q}_s$ be such that $Z_2'(\bar{q}_s) = 0$. Then $q(p) = \min\{\bar{q}_s, 1\}$ is the unique optimal solution to problem* (4.6).

Proof. We know $Z_2(\bar{q})$ is strictly concave because

$$Z_2''(\bar{q}) = \mathbb{E}_{\Lambda}\left[-\frac{2C\mu\Lambda^2}{(\mu - \bar{q}\Lambda)^3}\right] < 0$$

with $\bar{q} < \mu/\bar{\lambda}$. Given

$$Z_2'(\bar{q}) = \mathbb{E}_{\Lambda}\left[R\Lambda - \frac{C\mu\Lambda}{(\mu - \bar{q}\Lambda)^2}\right],$$

we have $Z_2'(0) \geq 0$ with the assumption that $R\mu \geq C$, and $Z_2'\left(\mu/\bar{\lambda} - \delta\right) < 0$ for some sufficiently small $\delta > 0$. Hence, there exists a $\bar{q}_s \in \left[0, \mu/\bar{\lambda}\right)$ such that $Z_2'(\bar{q}_s) = 0$. We thus conclude there must be one and only one optimal solution to (4.6), which is $q(p) = \min\{\bar{q}_s, 1\}$. □

We now again consider a profit maximizing firm that is allowed to set the value of p. Recall when the arrival rate is a known constant, the objective functions of profit maximization and social optimization are identical (see,

for example, [33]). The statement does not hold anymore in this case with a non-degenerate Λ. In fact, the expected profit rate is $\mathbb{E}_\Lambda[p \cdot q(p)\Lambda]$, where $p = \mathbb{E}_\Lambda[R - CW(q(p)\Lambda)]$. Hence, to maximize the profit rate, we need to solve

$$\max_{q(p)\in[0,1]\cap\left[0,\mu/\bar{\lambda}\right)} \mathbb{E}_\Lambda[q(p)\Lambda] \cdot \mathbb{E}_\Lambda[R - CW(q(p)\Lambda)]. \tag{4.7}$$

Let $Z_3(\bar{q}) := \mathbb{E}_\Lambda[\bar{q}\Lambda] \cdot \mathbb{E}_\Lambda[R - CW(\bar{q}\Lambda)]$ for $\bar{q} \in \left[0, \mu/\bar{\lambda}\right)$.

Proposition 4.3.2. *Let $\bar{q}_m$ be such that $Z_3'(\bar{q}_m) = 0$. Then $q(p) = \min\{\bar{q}_m, 1\}$ is the unique optimal solution to problem (4.7).*

Proof. Taking the first and and the second derivatives of $Z_3(\bar{q})$, we obtain

$$Z_3'(\bar{q}) = \mathbb{E}_\Lambda[\Lambda] \cdot \mathbb{E}_\Lambda\left[R - \frac{C\mu}{(\mu - \bar{q}\Lambda)^2}\right],$$

and

$$Z_3''(\bar{q}) = \mathbb{E}_\Lambda[\Lambda] \cdot \mathbb{E}_\Lambda\left[\frac{-2C\mu\Lambda}{(\mu - \bar{q}\Lambda)^3}\right].$$

With $Z_3'(0) \geq 0$ and $Z_3'\left(\mu/\bar{\lambda} - \delta\right) < 0$ for any $\delta > 0$ that is small enough, there must exist a $\bar{q}_m \in \left[0, \mu/\bar{\lambda}\right)$ such that $Z_3'(\bar{q}_m) = 0$. Since $Z_3(\bar{q})$ is strictly concave, the optimal solution to problem (4.7) is unique and given by $q(p) = \min\{\bar{q}_m, 1\}$. □

Theorem 4.3.3. *Let $\tilde{q}_s := \min\{\bar{q}_s, 1\}$ and $\tilde{q}_m := \min\{\bar{q}_m, 1\}$. Then, $\tilde{q}_s \leq \tilde{q}_m \leq \tilde{q}_e$.*

Proof. When $\bar{q}_s \leq 1$, $\tilde{q}_s = \bar{q}_s$, so $\bar{q}_s \leq \bar{q}_m$ leads to $\tilde{q}_s \leq \tilde{q}_m$, while when $\bar{q}_s > 1$, $\tilde{q}_s = 1$, and therefore $\tilde{q}_m = 1$ if $\bar{q}_s \leq \bar{q}_m$. Together with $\tilde{q}_e := \min\{\bar{q}_e, 1\}$, the argument indicates $\bar{q}_s \leq \bar{q}_m \leq \bar{q}_e$ is a sufficient condition for $\tilde{q}_s \leq \tilde{q}_m \leq \tilde{q}_e$.

As an increasing function of Λ, $W(\bar{q}_m\Lambda)^2$ is positively correlated with Λ (see Section 2 in [55]), and thus

$$\begin{aligned} Z_2'(\bar{q}_m) &= \mathbb{E}_\Lambda \left[R\Lambda - C\mu\Lambda \cdot W(\bar{q}_m\Lambda)^2 \right] \\ &\leq \mathbb{E}_\Lambda[\Lambda] \cdot \mathbb{E}_\Lambda \left[R - C\mu \cdot W(\bar{q}_m\Lambda)^2 \right] \\ &= Z_3'(\bar{q}_m) \\ &= 0 \end{aligned}$$

by Proposition 4.3.2. Recall $Z_2'(\bar{q}_s) = 0$ by Proposition 4.3.1, which implies that $\bar{q}_s \leq \bar{q}_m$ given the strict concavity of $Z_2(\cdot)$.

We now consider the relationship between $\bar{q}_m$ and $\bar{q}_e$. According to (4.5), $R\mu = C$ if $\bar{q}_e = 0$. Assume $\bar{q}_m > 0$ in this case, and we have

$$Z_3'(\bar{q}_m) = \mathbb{E}_\Lambda[\Lambda] \cdot \mathbb{E}_\Lambda \left[R - \frac{C\mu}{(\mu - \bar{q}_m\Lambda)^2} \right] < 0,$$

which violates $Z_3'(\bar{q}_m) = 0$. Hence, $\bar{q}_m = \bar{q}_e = 0$. We also claim $\bar{q}_m \leq \bar{q}_e$ if $\bar{q}_e > 0$. Otherwise given $Z_3(0) = 0$,

$$\frac{\bar{q}_m - \bar{q}_e}{\bar{q}_m} \cdot Z_3(0) + \frac{\bar{q}_e}{\bar{q}_m} \cdot Z_3(\bar{q}_m) \geq 0 = Z_3(\bar{q}_e),$$

which contradicts the strict concavity of $Z_3(\cdot)$. □

Theorem 4.3.3 suggests that both $\tilde{q}_s$ and $\tilde{q}_m$ can be induced by enforcing appropriate entering fees, while the profit maximizer in fact prefers a lower p. We also notice that $Z_2(\bar{q}) \leq Z_3(\bar{q})$ since $\bar{q}\Lambda$ and $W(\bar{q}\Lambda)$ are positively correlated

as non-decreasing functions of Λ. More precisely, we derive

$$\begin{aligned}
\mathrm{Cov}[\bar{q}\Lambda, W(\bar{q}\Lambda)] &= -\mathrm{Cov}\left[\mu - \bar{q}\Lambda, \frac{1}{\mu - \bar{q}\Lambda}\right] \\
&= \mathbb{E}_\Lambda[\mu - \bar{q}\Lambda]\mathbb{E}_\Lambda\left[\frac{1}{\mu - \bar{q}\Lambda}\right] - \mathbb{E}_\Lambda[1] \\
&\geq \mathbb{E}_\Lambda[\mu - \bar{q}\Lambda] \cdot \frac{1}{\mathbb{E}_\Lambda[\mu - \bar{q}\Lambda]} - 1 \\
&= 0
\end{aligned}$$

by Jensen's inequality, where the equality holds if and only if $\bar{q} = 0$. That is, the aggregate consumer surplus is negative if a profit maximizer charges a non-zero entrance fee. Furthermore, with $Z_2(\bar{q}_e) < Z_3(\bar{q}_e) = 0$ for $\bar{q}_e \neq 0$, we find social welfare is typically negative if no entering fee is enforced to suppress customers' willingness to join.

4.4 Computational Results

Let $\Omega = \{\lambda_1, \ldots, \lambda_d\}$ denote the state space of Λ, and

$$\mathbf{prob} := \left(\mathbb{P}\left(\Lambda = \lambda_1\right), \ldots, \mathbb{P}\left(\Lambda = \lambda_d\right)\right).$$

In Table 4.1 and Table 4.2, we present numerical results for observable and unobservable systems respectively with $\mu = 1$, $R = 12$ and $C = 1$. We assume a customer joins in the threshold case where his individual profit is zero, and we denote the amount of the entering fee associated with social optimization and profit maximization as $\tilde{p}_s$ and $\tilde{p}_m$ respectively.

The result agrees to the findings in Section 4.2 and Section 4.3, and it shows how decisions vary with different perspectives.

Table 4.1: Joining and Pricing Strategies for Observable Queues

	$\tilde{n}_e$	$\tilde{n}_s$	$\tilde{n}_m$	$\tilde{p}_s$	$\tilde{p}_m$
$\Omega = \{0.5\}, \mathbf{prob} = (1.0)$	12	6	2	$(5, 6]$	10
$\Omega = \{1\}, \mathbf{prob} = (1.0)$	12	4	3	$(7, 8]$	9
$\Omega = \{2\}, \mathbf{prob} = (1.0)$	12	3	2	$(8, 9]$	10
$\Omega = \{0.5, 2\}, \mathbf{prob} = (0.5, 0.5)$	12	6	2	$(5, 6]$	10
$\Omega = \{0.5, 1, 2\}, \mathbf{prob} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	12	5	2	$(6, 7]$	10
$\Omega = \{0.5, 1, 2\}, \mathbf{prob} = (0.2, 0.3, 0.5)$	12	4	2	$(7, 8]$	10
$\Omega = \{0.5, 1, 2, 5\}, \mathbf{prob} = (0.2, 0.3, 0.4, 0.1)$	12	4	2	$(7, 8]$	10

Table 4.2: Joining and Pricing Strategies for Unobservable Queues

	$\tilde{q}_e$	$\tilde{q}_s$	$\tilde{q}_m$	$\tilde{p}_s$	$\tilde{p}_m$
$\Omega = \{0.5\}, \mathbf{prob} = (1.0)$	1	1	1	10	10
$\Omega = \{1\}, \mathbf{prob} = (1.0)$	0.917	0.711	0.711	8.54	8.54
$\Omega = \{2\}, \mathbf{prob} = (1.0)$	0.458	0.356	0.356	8.53	8.53
$\Omega = \{0.5, 2\}, \mathbf{prob} = (0.5, 0.5)$	0.478	0.369	0.394	9.48	9.02
$\Omega = \{0.5, 1, 2\}, \mathbf{prob} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	0.485	0.386	0.411	9.58	9.14
$\Omega = \{0.5, 1, 2\}, \mathbf{prob} = (0.2, 0.3, 0.5)$	0.478	0.375	0.393	9.27	8.92
$\Omega = \{0.5, 1, 2, 5\}$, $\mathbf{prob} = (0.2, 0.3, 0.4, 0.1)$	0.198	0.167	0.180	10.21	9.79

Chapter 5

Integrated Replenishment and Inbound Transportation with a Location-Based Model

5.1 Introduction

The chapter is motivated by a project with an engine assembly plant in Texas, where the goal is to help reduce the total inventory holding and inbound transportation cost of parts used in the assembly line in the plant. We are specifically interested in procured parts and subassemblies that go straight into final assembly, which are typically expensive and heavy, and are shipped directly from about one hundred suppliers to the plant. As a newly moved factory, the assembly plant's contract suppliers are mostly located hundreds of miles away. Therefore, wrongly selected freight shipping modes or frequencies for parts may result in an astronomical logistics cost, while appropriate ones can lead to substantial savings. These are the decisions we are concerned about.

We consider two inbound transportation modes: full truckload (FTL) shipping and less than truckload (LTL) shipping. The FTL trucks are dedicated to the plant, and each of them serves a predetermined group of suppliers with a fixed cost per trip. On the contrary, the LTL shipping company does

not accommodate customized routing, so the cost per trip is computed for each supplier separately. Due to an attempt to constrain operational complexity, such as the amount of paperwork at receiving docks, we only allow a finite number of possible order frequencies, and we do not split shipments from the same supplier across different trucks. In addition, all suppliers belonging to an FTL group must have the same order frequency and be visited together according to it.

We study an integrated problem here because a sequential ordering and shipping decision is often sub-optimal. In fact, we are likely to obtain long order cycle times and a small total number of shipments when minimizing the inbound transportation cost and exactly the opposite when optimizing the inventory holding cost. For the sake of simplicity, we pursue static strategies that are based on stable demand assumed to be calculated accurately for a year, where "static" means we make decisions once and then follow them to replenish parts periodically. A major element of the problem is FTL routing, which can be formulated as a vehicle routing problem (VRP). However, doing so greatly complicates the model we need to solve for gaining the insights especially given the number of suppliers in the system, and we instead apply a location-based model that approximates the exact routing cost with an upper bound.

We believe our findings can be generalized to any plant that assembles complex equipment worth hundreds of thousands of dollars and possibly made of heavy metals, which include but not limited to engines, medical equipment

and semiconductor equipment. The discussion can also shed some light on the debate of production outsourcing.

In the rest of the chapter, we present an overview of the literature relevant to our work in Section 5.2 and the upper-bound model to find a heuristic solution in Section 5.3. We then show results of numerical experiments in Section 5.4.

5.2 Literature Review

The literature on when and how many to order goes back to Harris [30] who derives the economic order quantity, and the problem has evolved over time with additional concerns. For example, Baumol and Vinod [11] incorporate a transportation cost that is linear on the shipment size into their analysis for a single commodity, and they refer to it as the inventory-theoretic model for freight shipment decisions. Winston [59] includes an overview of the early development of the idea and concludes that endogenizing inventory-based decisions is vital in freight transportation problems. A more recent review is covered by Min and Zhou [44] from the perspective of supply chain modeling, as well as Meixell and Norbis [43] from the angle of mode and carrier selection.

For systems with multiple suppliers providing heterogeneous products, the structure of shipping costs for each commodity is generally complex with routing options. In fact, the integrated ordering and shipping problem in this case is an extension of the VRP, of which the main algorithms are summarized by Laporte [38]. As for the joint problem, there is a substantial literature on

outbound shipping while not as much on inbound transportation. However, as pointed out by Andersson et al. [3], it is easy to transform the topology of one problem into the other. Hence, we survey related work in both scenarios.

The review by Sarmiento and Nagi [52] divides the work on integrated inventory control and outbound routing into two classes: one on the inventory routing problem, such as in Dror and Ball [24] and Chien et al. [21], and the other on the "one-depot, multiple-retailers" problem. Our work is most related to those having a "distribution-inventory" structure along the latter direction. For example, Federgruen and Zipkin [26] are believed to be the first to formulate a single model that combines inventory allocation and vehicle routing, which can be solved by a modified interchange heuristic or an exact algorithm based on generalized Benders' decomposition. Burns et al. [17] use the density of customers as the input instead of the specific location of each to compare the distribution costs incurred by direct shipping and peddling. The same approximation technique is also used by Daganzo [23] for an inbound shipping problem. In addition, Anily and Federgruen [4] develop accurate bounds for the long-run average total cost and an optimal heuristic in an asymptotic sense for systems with a single product. Viswanathan and Mathur [56] consider a similar problem for distribution networks with multiple products, and they design a heuristic to find a stationary nested joint replenishment policy. There are other heuristic algorithms in this field, such as the ones proposed by Bramel and Simchi-Levi [16] and Sindhuchao et al. [53]. In particular, we adopt the idea of location-based modeling for the capacitated VRP presented

in the former paper, which is motivated by the earlier work in Bramel et al. [15]. The method also appears in a later review by Bertsimas and Simchi-Levi [12]. For systems with stochastic demand, Qu et al. [49] decompose a long-run cost minimization model into an inventory master problem and a transportation subproblem, and they obtain solutions with good performance by solving the two parts iteratively. Finally, there is a recent review by Coelho et al. [22] that includes several papers mentioned above.

Another element of the problem we have is the selection between FTL and LTL services, which requires cost estimates for both. Note that FTL shipping is also referred to as truckload shipping or TL shipping in the literature, of which the price is basically computed on a per-mile basis (see Swenseth and Godfrey [54]). On the other hand, the cost structure of LTL shipping is complex (see Rieksts and Ventura [51]) and not modeled analytically until the recent study by Özkaya et al. [48]. We apply their results to our problem, and the details are provided in Section 5.3.

5.3 An Upper-Bound Model

Recall our objective is to minimize the annual cost of inbound shipping and inventory holding and explore the associated insights. The assembly plant is open five days a week, so a year can have 260-262 working days in total. We pick the middle number and assume there are always 261 of them per year to make static decisions. We also assume a zero inventory level before the first order arrives. In addition, we imagine a "standard virtual" part for each

supplier as the representative of all provided. The standard part has an average unit weight and purchase price weighted by the individual part demand, and its demand equals to the total demand of goods from the supplier. We solve the planning problem for the standard parts to obtain the supplier-level results, which can be translated into part-level decisions by allocating order quantities proportionally. The simplification is reasonable and helpful in that parts from a supplier tend to be different models of a product with similar characteristics. Hence, compared to a part-level formulation, the supplier-level model is more robust to demand fluctuations caused by dynamic production schedules in practice.

Let $\mathcal{K}$ denote the set of all possible order frequencies, and we define τ_k as the corresponding order cycle time for any $k \in \mathcal{K}$. Given supplier j and order frequency k, we know the part order quantity $q_{jk} = \frac{d_j \tau_k}{261}$, where d_j is the annual demand for the associated standard part. Let h_j be the product of a fixed annual interest rate and the purchase price of a single part. We can write the annual inventory holding cost incurred by having supplier j as $\frac{1}{2} h_j q_{jk}$.

Based on the regression result for LTL pricing in Özkaya et al. [48], for each supplier, we are able to derive the best LTL shipping frequency analytically. Among the eight predictors suggested by the authors, which are also listed and defined in Table 5.1, Mile, Freight Class, Freight Index and Origin Region are fixed for a given supplier, Shipper Index and Destination Region solely rely on characteristics of the assembly plant, Carrier Type is set by preference beforehand, and Weight is the only factor left to be determined.

Being a linear function of Weight, the part order quantity is thus the only argument in the fitted LTL cost function provided in [48], and the function is in fact quadratic. Note the value obtained from the fitted model is the base price that does not include the fuel surcharge, which is typically a certain percentage of the base price in the LTL case. Besides, there is a weight limit for any LTL shipment. Taking all these into account, we find the minimum annual shipping and inventory holding cost for serving supplier j by LTL is

$$c_j = \min_{k \in \mathcal{K}} \left\{ A_j \left(q_{jk} \right) + \frac{1}{2} h_j q_{jk} \right\},$$

where $A_j(q_{jk})$ represents the sum of the base price and the fuel surcharge for shipping q_{jk} parts from supplier j, and it is arbitrarily large if the shipment weight exceeds the weight limit. We can solve for c_j and the corresponding shipping frequency analytically or just by enumeration and comparison. Let z_j be 1 if supplier j is indeed served by LTL and 0 otherwise. The annual total cost associated with LTL services is then $\sum_{j \in \mathcal{J}} c_j z_j$ with $\mathcal{J}$ being the set of all suppliers.

We now consider FTL services. Both the base price and the fuel surcharge for an FTL trip are on a cost per mile basis except there exists a minimum base price for short-distance trips. Recall we want to build a model that is easier to solve than one based on VRP, so we approximate the routing cost and transform the exact problem into a capacitated facility location problem (CFLP). We implement the idea of "star connection" (see [16]) to ensure the optimal objective value obtained is an upper bound for the minimum total cost in practice. That is to say, whenever parts are ordered from a

Table 5.1: Predictors in the Regression Model for LTL Pricing

Predictor	Definition
Weight	Total shipment weight in pounds
Mile	One-way distance between the origin and the destination in miles
Freight Class	Contracted freight class that is mainly determined by the contracted freight density
Origin Region	Region the origin belongs to, such as Mid West, South Central and so forth
Destination Region	Region the destination belongs to, such as Mid West, South Central and so forth
Carrier Type	National or regional depending on how many states are served by the carrier
Shipper Index	Score that reflects the shipper's negotiation power in the market
Freight Index	Score computed with the actual freight class that is primarily related to the actual freight density

supplier served by an FTL route, we assume an empty FTL truck is sent from the assembly plant to the "seed" of the route. The truck then takes a round trip between the seed and each supplier on the route to load parts at given quantities before it drives back to the plant. By doing so, we can refer to a route by its seed (say, $i \in \mathcal{I}$) and its frequency ($k \in \mathcal{K}$). Imagining potential routes as potential facilities in CFLP, we aim to decide which route to activate and which suppliers it serves. Let f_{ik} denote the annual FTL cost of getting to seed i and coming back to the plant with frequency k, and let s_{ijk} be the

annual cost of including supplier j into the route represented by seed i and frequency k due to the additional traveling distance. Note that we only take the minimum base price into consideration when computing f_{ik}. Also, the annual cost equals to the cost per trip times the number of trips in a year, which is given by $\frac{261}{\tau_k}$. We define y_{ik} as the binary variable representing whether the route with seed i and frequency k is activated and x_{ijk} as the binary variable denoting whether supplier j is assigned to the route. The total FTL shipping cost is then

$$\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} f_{ik}y_{ik} + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{K}} s_{ijk}x_{ijk},$$

and the annual inventory holding cost for parts shipped by FTL is

$$\frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{K}} h_j q_{jk} x_{ijk}.$$

We incorporate the LTL option and formulate the entire problem with the notation summarized in Table 5.2 as follows:

$$\text{Minimize } \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} f_{ik}y_{ik} + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{K}} s_{ijk}x_{ijk} + \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{K}} h_j q_{jk} x_{ijk} + \sum_{j\in\mathcal{I}} c_j z_j$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} x_{ijk} + z_j = 1 \qquad \forall j \in \mathcal{I}, \tag{5.1}$$

$$x_{ijk} \le y_{ik} \qquad \forall i \in \mathcal{I}, j \in \mathcal{I}, k \in \mathcal{K}, \tag{5.2}$$

$$\sum_{j\in\mathcal{I}} w_j q_{jk} x_{ijk} \le g_{ik} y_{ik} \quad \forall i \in \mathcal{I}, k \in \mathcal{K}, \tag{5.3}$$

$$x_{ijk}, y_{ik}, z_j \in \{0,1\} \quad \forall i \in \mathcal{I}, j \in \mathcal{I}, k \in \mathcal{K}. \tag{5.4}$$

According to (5.1), all suppliers need to be served by either an FTL route or

Table 5.2: Notation for the Upper-Bound Model

Sets and Indices	
$\mathcal{I}$	set of suppliers
$\mathcal{K}$	set of possible order frequencies
$i \in \mathcal{I}$	index of FTL route seeds
$j \in \mathcal{I}$	index of suppliers
$k \in \mathcal{K}$	index of order frequencies
Input Parameters	
q_{jk}	order quantity for supplier j with frequency k
w_j	weight per part from supplier j
g_{ik}	capacity of the FTL truck running on the route with seed i and frequency k
f_{ik}	annual cost for activating the FTL route with seed i and frequency k
s_{ijk}	annual cost for assigning supplier j to the FTL route with seed i and frequency k
h_j	annual inventory holding cost per part from supplier j
c_j	minimum annual shipping and inventory holding cost incurred by serving supplier j with an LTL carrier
Decision Variables	
y_{ik}	binary variable which equals 1 if the FTL route with seed i and frequency k is activated, 0 otherwise
x_{ijk}	binary variable which equals 1 if supplier j is assigned to the FTL route with seed i and frequency k, 0 otherwise
z_j	binary variable which equals 1 if supplier j is served by an LTL carrier, 0 otherwise

an LTL carrier. Constraint (5.2) indicates a supplier can be assigned to a route only if it is activated. Finally, (5.3) is the capacity constraint for FTL trucks, and we measure the capacity only by weight because of the high freight density or the low freight class. All the constraints are standard for CFLP except we add z_j to allow LTL shipping.

An optimal solution to the formulation provides a feasible ordering and shipping plan, and we are at least able to lower the FTL shipping cost with better routing than the star connection. Hence, the model is indeed an upper-bound model. In fact, we can compute the real routing cost for a given solution by solving the traveling salesman problem (TSP) for each activated FTL route. The procedure typically requires low computational effort due to the small number of suppliers covered by a route.

5.4 Computational Results

We generate parameters to mimic the real data collected for the engine assembly plant.

We consider an instance with 100 suppliers in total, whose Cartesian coordinates are generated from two independent normal distributions with mean 0 and standard deviation 200 (miles). The assembly plant is assumed located at the point $(-\gamma, -\gamma)$, where $\gamma = 600$ by default. We measure all distances here using the Euclidean metric, while we recommend replacing them with driving distances when possible.

The daily demand for each standard part is an integer randomly selected between 10 and 500. The unit weight (pounds) of a part is generated from the uniform distribution $\mathcal{U}(2, 50)$, and its unit purchase price ($) is $\mathcal{U}(2, 30)$ multiples of its unit weight.

We assume τ_k is an element of the set $\{1, 2, 3, 4, 5, 10, 15, 20\}$, which means a part is ordered every day, every 2-4 days, every week or every 2-4 weeks, with $\tau_1 = 1$ without loss of generality. An annual interest rate of 25% is used to compute the inventory holding cost. For FTL shipping, we set g_{ik} to be identical for all i and k following the literature, and the default value of g_{ik} is $35,000$ pounds, while an LTL shipment cannot weigh more than $15,000$ pounds. Recall we do not allow splitting of orders in any form. Hence, LTL services are not available for supplier j if $w_j q_{j1} > 15,000$. It is rare to have any supplier with daily shipment weight over the FTL truck capacity, and we do not encounter such suppliers in the instance, but if we do, we can just take the supplier out and solve the optimization model for the rest suppliers. Otherwise, the problem is infeasible.

We then generate LTL and FTL shipping costs. We select the LTL carrier type to be national considering the locations of the assembly plant and the contract suppliers. The destination region is South Central, while origin regions depend on supplier locations. To simplify the analysis, we use a single origin region of Mid West since it is regarded as "the most industrialized region" (see [48]) and most suppliers we have are indeed there. Assuming a uniform freight class of 70 across all suppliers, we approximate the corresponding freight

index with equation (4) in [48], which yields the result of 49.63. Note that numerical predictors are standardized with coefficients provided in Table 5 of [48], so we normalize the shipper index to zero to represent an average case without further information. Together with supplier-specific data on the shipping distance and the shipment weight, we obtain the LTL base price, and the LTL fuel surcharge is 15% of the base price. We calibrate the parameters by evaluating some old instances and comparing the fitted results with the real carrier quotes.

In the FTL case, the cost per trip for a truck is \$500 in total or \$1 per mile, whichever is larger, plus the FTL fuel surcharge of \$0.35 per mile. We denote the one-way distance (in miles) between the assembly plant and node j by Δ_j, and the one-way distance (in miles) between nodes i and j by $\delta_i j$. We can thus compute f_{ik} and s_{ijk} as

$$f_{ik} = (\max\{\Delta_i \times 2 \times \$1, \$500\} + \Delta_i \times 2 \times \$0.35) \times \frac{261}{\tau_k},$$

and

$$s_{ijk} = (\delta_{ij} \times 2 \times \$1 + \delta_{ij} \times 2 \times \$0.35) \times \frac{261}{\tau_k}.$$

We show representative cost curves for FTL and LTL shipping in Figure 5.1 and Figure 5.2. Although the LTL curve is below the FTL curve in both charts, LTL services can be more expensive than FTL services in terms of cost per pound.

We implement the upper-bound model in GAMS, which can be solved to optimality within a minute by CPLEX. We present numerical results of

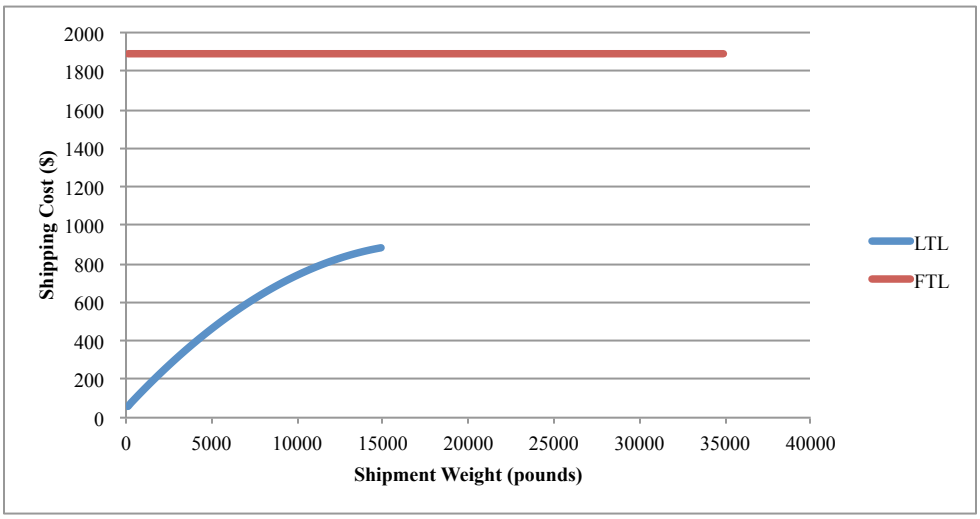


Figure 5.1: The chart depicts the LTL and FTL shipping costs with a one-way shipping distance of 700 miles.

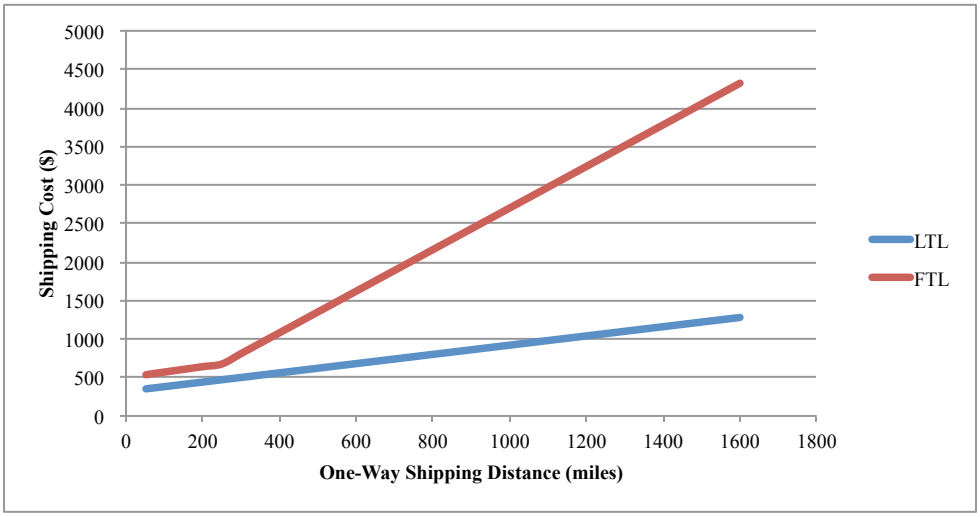


Figure 5.2: The chart depicts the LTL and FTL shipping costs with shipment weight of 10,000 pounds.

the upper-bound model for different parameter settings in Table 5.3, where "Benchmark" means all parameters are kept at their default values, while the others have only one group of parameters changed as stated in the first column. We report the following measures evaluated at the optimal solution to the upper-bound model: the annual FTL shipping cost estimated by the model (ftl_ub), the annual FTL shipping cost in practice by solving the TSP (ftl_tsp), the annual inventory holding cost associated with FTL shipping (ftl_inv), the annual shipping and inventory holding cost associated with LTL services (ltl_tot), the total number of FTL routes (#rou) and the total number of suppliers served by FTL shipping (#sup). The locations of suppliers as well as the characteristics of their products remain the same in all cases. Note the cost figures are in thousands of dollars, and the adjusted demand rate is rounded to the nearest integer.

It is indicated by [16] that the real optimal routing cost is closely approximated by the star connection heuristic when the number of "customers" in a compact region is large enough, where customers are suppliers in our case. We compare the first two columns of Table 5.3 and obtain an average relative error of 5.25%, which suggests the upper-bound model performs well given the current supplier density. We find the approximation is more accurate when γ is large because the traveling distance due to local routing takes a smaller percentage of the total shipping distance. In addition, the approximation works better with low demand rates and a low interest rate. We believe the reason can be intuitively explained as follows. There are two ways of saving in the

Table 5.3: Results for the Upper-Bound Model with Parameter Changes

	ftl_ub	ftl_tsp	ftl_inv	ltl_tot	#rou	#sup
Benchmark	8,238	7,956	1,044	6,187	21	55
$\gamma = 0$	4,815	4,306	2,031	799	35	82
$\gamma = 1200$	9,573	9,267	471	15,190	8	23
125% fuel surcharges	7,838	7,562	822	7,507	18	49
150% fuel surcharges	7,837	7,598	752	8,244	16	42
75% demand rates	4,716	4,542	862	6,347	17	47
125% demand rates	12,526	11,413	1,528	4,834	26	69
75% interest rate	8,421	8,193	866	5,669	23	57
125% interest rate	8,217	7,794	1,139	6,587	19	55

problem: consolidation across time and consolidation across space. The former is possible for both FTL and LTL shipping, while the latter is specific to FTL services. When we experience high demand rates, we have limited consolidation opportunities, which makes grouping suppliers located far from each other an attractive option. In the case of a high interest rate, we shift our focus to consolidation across space and tend to assign more suppliers to a route since consolidation across time results in a high inventory holding cost. Under such scenarios, the difference between the star-connection distance and the real routing distance is relatively large, and so is the difference between "ftl_ub" and "ftl_tsp".

We then observe the first three rows of Table 5.3. It is interesting to see that LTL services are preferred to FTL services in long-distance shipping. The

result is not surprising because the FTL cost is more sensitive to the shipping distance as shown in Figure 5.2, but it suggests a major policy change for the assembly plant that motivates our study. The plant moved to Texas from the center of the contract suppliers, and the managers are questioning whether they should keep the original daily FTL shipping policy especially when LTL prices seem competitive.

The fuel surcharges used in the benchmark case correspond to low fuel prices, and we see LTL shipping is less affected by a rise of fuel surcharges. Besides, part of the reason why we rely more on FTL services with high demand rates is due to the weight limit on each LTL shipment. Finally, we observe the third column of Table 5.3 and notice the inventory holding cost is significantly less than the shipping cost in general, and hence the benchmark case and the last two experiments on the interest rate yield similar results.

We solve several variants of the integrated problem and report the results in Table 5.4. Our purpose is to separate the ordering and shipping decisions so that we understand what we gain by considering them simultaneously. All parameters are now at the default values, so "Benchmark" refers to exactly the same instance in Table 5.3 and Table 5.4. We first minimize the inventory holding cost by ordering all the parts daily, and we use only FTL shipping (daily_ftl) or the best shipping mode (daily_best) for transportation. We then ignore the inventory holding cost in the upper-bound model and optimize the resulting inbound shipping problem (no_inv). In addition to "ftl_ub", we present the annual LTL shipping cost (ltl_ub), the annual

Table 5.4: Results for Variants of the Integrated Problem

	ftl_ub	ltl_ub	inv	tot
Benchmark	8,238	5,232	1,999	15,470
daily_ftl	17,096	0	1,255	18,351
daily_best	8,791	6,334	1,255	16,381
no_inv	9,793	3,194	3,387	16,373

inventory holding cost (inv) and the annual total cost (tot) incurred by the optimal decisions in thousands of dollars.

Table 5.4 indicates we can save more than $900,000 per year by solving the integrated problem. Also, it demonstrates that LTL shipping is an important option to include when long-distance freight transportation is required, which is counterintuitive for such assembly plants.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions and Contributions

Resource allocation can refer to a variety of problems depending on the application area, and we concentrate on staffing and pricing of service systems as well as ordering and shipping in logistics systems.

First, we extend previous work on staffing a large-scale service system with exponential service times and Poisson arrivals in the presence of arrival-rate uncertainty. In particular, we assume only the support and mean of the arrival-rate distribution have been estimated. The objective is to minimize the staffing cost without jeopardizing the service quality in a long-run average sense, where the staffing level needs to be determined before the arrival rate is realized. We measure the service quality by the probability that a customer has to wait for services in the queue, and a QoS constraint is imposed. The JVLZ upper bound is applied to computing the QoS metric instead of the exact Erlang-C formula because it is numerically unstable in large-scale systems. Depending on how "nature" chooses the actual arrival-rate distribution, we formulate a meta-distributed model and also an adversarial model. For both models, we develop the two-step approximation techniques, as in the case

with a given arrival-rate distribution, to provide tractable solutions to the associated staffing problems. That is, we first identify the key scenario that always results in a non-trivial QoS level as the system size scales up. The staffing problem then reduces to one with a deterministic arrival rate decided by the key scenario, and we solve it as a constraint satisfaction problem. We prove the solutions obtained are asymptotically optimal in the context of the classic Halfin-Whitt regime. The results are verified by various numerical experiments, and the value of information is quantified to demonstrate how many agents are added due to incomplete distributional knowledge.

We next consider the staffing problem in a multi-station system that handles different service requests with dedicated server pools. We show the optimal staffing strategy is unique when the arrival rates of all stations are deterministic or in a two-station case with stochastic arrival rates generated by a discrete bivariate distribution.

From a game-theoretic perspective, we expand the joining or balking analysis for $M/M/1$ queues by incorporating parameter uncertainty. Assuming the arrival rate is randomly selected from a given distribution, we derive the optimal joining threshold when queue lengths are observable as well as the optimal joining percentage when queue lengths are unobservable for an individual customer, the social optimizer and the profit maximizer. In the observable case, we prove the inequality from Naor's work still holds when the arrival rate is stochastic. That is to say, individual customers are more willing to join the queue than they should as a group, and the social optimizer can

resolve the issue by authorizing the system to impose an appropriate entering fee. However, a profit maximizing firm endeavors to keep fewer customers in the system and thus charges a higher price than desired by the social optimizer. The former part of the statement remains true in unobservable queues, while the latter part is reversed. In fact, the profit maximizer prefers a joining percentage that is between the ones chosen by an individual customer and the social optimizer respectively. In other words, the optimal strategies for the social optimizer and the profit maximizer are not aligned as with a deterministic arrival rate. We also notice the aggregate consumer surplus can be negative in expectation with arrival-rate uncertainty.

Finally, we formulate a location-based model for an integrated ordering and inbound shipping problem, where the goal is to minimize the total cost of inventory holding, FTL routing and LTL shipping by selecting proper order frequencies and shipping methods. Although the work is motivated by an engine assembly plant, the results can be applied to any manufacturer that regularly purchases heavy and expensive parts from a large number of suppliers. The location-based model is not an exact formulation of the problem since the FTL routing cost is approximated, but given the typical problem size, our model serves better for exploring the insights with minimum computational efforts required to solve to optimality. We adopt the star connection heuristic so that the optimal value obtained is guaranteed to be an upper bound of the optimal total cost in practice. We evaluate the real routing cost by solving a TSP for suppliers assigned to the same route. Numerical experiments suggest

the model performs well in approximating the routing cost for typical parameter settings, and we conclude from the optimal results that LTL shipping deserves more attention for long-distance freight transportation.

6.2 Future Research

Last but not least, we discuss possible research directions that are related to our work in this dissertation.

For the staffing problems studied in Chapter 2 and Chapter 3, we assume the arrival-rate distribution is discrete. If the distribution is instead continuous, the key scenario concept used in this dissertation is no longer well defined. In fact, there might not exist an asymptotically optimal staffing level with the current scaling mechanism. Extending the analysis to handle continuous arrival-rate distributions may also help interpret the value of information result, because it is expected to eliminate the zig-zag pattern that occurs when the key scenario changes in the discrete case.

The multi-station staffing problem in Chapter 3 lies at the interface of queueing analysis and nonlinear optimization. Numerical experiments on systems with few customer classes suggest there is a single optimal way of assigning agents to stations even if we replace the Erlang-C formula with the JVLZ upper bound. The property can shed light on how to solve the problem if it is proved true. Following the framework in Chapter 2, we can then proceed with additional levels of stochasticity. The problem has applications in not only service optimization but inventory management for make-to-order

manufacturers as well. For example, we are intrigued to know the best static plan of ordering common components and allocating them to products.

Other than joining and balking, customers arriving at an observable queue may prefer waiting outside of the system with a lower cost. The problem has been formulated as a Markov decision process in the literature for the case where only one customer is strategic and all others join without thinking (see [41]), while it is still an open question to characterize equilibria when all customers are "smart". A natural further step is to investigate a version where the arrival rate is random. Unlike the observable model in Chapter 4, customers are now concerned about the arrival rate when choosing actions, so a larger impact of the parameter uncertainty on equilibrium states is anticipated.

For the logistics problem in Chapter 5, theoretical support for the numerical findings is desired. Also, we are interested in providing a tight lower bound of the optimal total cost as a benchmark for the upper-bound result. Moreover, a potential obstacle in translating supplier-level decisions to part-level decisions is the involvement of package quantities, which affects the flexibility in mixing and matching products. To resolve the issue, either a part-level dynamic lot-sizing model or a set of practical adjustment rules can be of great use.

Bibliography

[1] P. Afèche and B. Ata. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management*, 15(2):292–304, 2013.

[2] Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[3] H. Andersson, A. Hoff, M. Christiansen, G. Hasle, and A. Løkketangen. Industrial aspects and literature survey: Combined inventory management and routing. *Computers & Operations Research*, 37(9):1515–1536, 2010.

[4] S. Anily and A. Federgruen. One warehouse multiple retailer systems with vehicle routing costs. *Management Science*, 36(1):92–114, 1990.

[5] A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

[6] C. Bandi, D. Bertsimas, and N. Youssef. Robust queueing theory. *Operations Research*, 63(3):676–700, 2015.

[7] O. Baron and J. Milner. Staffing to maximize profit for call centers with alternate service-level agreements. *Operations Research*, 57(3):685–700, 2009.

[8] A. Bassamboo, J. M. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research*, 54(3):419–435, 2006.

[9] A. Bassamboo, R. S. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686, 2010.

[10] A. Bassamboo and A. Zeevi. On a data-driven method for staffing large call centers. *Operations Research*, 57(3):714–726, 2009.

[11] W. J. Baumol and H. D. Vinod. An inventory theoretic model of freight transport demand. *Management Science*, 16(7):413–421, 1970.

[12] D. J. Bertsimas and D. Simchi-Levi. A new generation of vehicle routing research: Robust algorithms, addressing uncertainty. *Operations Research*, 44(2):286–304, 1996.

[13] O. Besbes and C. Maglaras. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research*, 57(6):1438–1450, 2009.

[14] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[15] J. Bramel, E. G. Coffman, P. W. Shor, and D. Simchi-Levi. Probabilistic analysis of the capacitated vehicle routing problem with unsplit demands. *Operations Research*, 40(6):1095–1106, 1992.

[16] J. Bramel and D. Simchi-Levi. A location based heuristic for general routing problems. *Operations Research*, 43(4):649–660, 1995.

[17] L. D. Burns, R. W. Hall, D. E. Blumenfeld, and C. F. Daganzo. Distribution strategies that minimize transportation and inventory costs. *Operations Research*, 33(3):469–490, 1985.

[18] J. Cardoso, H. Fromm, S. Nickel, G. Satzger, R. Studer, and C. Weinhardt, editors. *Fundamentals of service systems*. Springer, 2015.

[19] B. P. K. Chen and S. G. Henderson. Two issues in setting call center staffing levels. *Annals of Operations Research*, 108(1):175–192, 2001.

[20] H. Chen and M. Z. Frank. State dependent pricing with a queue. *IIE Transactions*, 33(10):847–860, 2001.

[21] T. W. Chien, A. Balakrishnan, and R. T. Wong. An integrated inventory allocation and vehicle routing problem. *Transportation Science*, 23(2):67–76, 1989.

[22] L. C. Coelho, J. F. Cordeau, and G. Laporte. Thirty years of inventory routing. *Transportation Science*, 48(1):1–19, 2014.

[23] C. F. Daganzo. Supplying a single location from heterogeneous sources. *Transportation Research B*, 19(5):409–419, 1985.

[24] M. Dror and M. Ball. Inventory/routing: Reduction from an annual to a short-period problem. *Naval Research Logistics*, 34:891–905, 1987.

[25] N. M. Edelson and D. K. Hildebrand. Congestion tolls for poisson queueing processes. *Econometrica*, 43:81–92, 1975.

[26] A. Federgruen and P. Zipkin. A combined vehicle routing and inventory allocation problem. *Operations Research*, 32(5):1019–1037, 1984.

[27] N. Gans, H. Shen, Y. Zhou, K. Korolev, A. McCord, and H. Ristock. Parametric stochastic programming models for call-center workforce scheduling. *Working paper*, 2009.

[28] I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115, 2010.

[29] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[30] F. W. Harris. How many parts to make at once. *Operations Research*, 38:947–950, 1990 (reprinted from *Factory, The Magazine of Management*, 10:135–136, 152, 1913).

[31] J. M. Harrison and A. Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1):20–36, 2005.

[32] R. Hassin. On the optimality of first come last served queues. *Econometrica*, 53:201–202, 1985.

[33] R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media, 2003.

[34] M. Haviv and B. S. Randhawa. Pricing in queues without demand information. *Manufacturing & Service Operations Management*, 16(3):401–411, 2014.

[35] A. A. Jagers and E. A. van Doorn. Convexity of functions which are generalizations of the erlang loss function and the erlang delay function. *SIAM Review*, 33(2):281–282, 1991.

[36] A. J. E. M. Janssen, J. S. H. Van Leeuwaarden, and B. Zwart. Refining square root safety staffing by expanding Erlang C. *Operations Research*, 59(6):1512–1522, 2011.

[37] Y. L. Koçağa, M. Armony, and A. R. Ward. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, 2015.

[38] G. Laporte. The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(3):345–358, 1992.

[39] S. Liao, C. Van Delft, and J. P. Vial. Distributionally robust workforce scheduling in call centres with uncertain arrival rates. *Optimization Methods and Software*, 28(3):501–522, 2013.

[40] S. Liao, G. Koole, C. Van Delft, and O. Jouini. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum*, 34(3):691–721, 2012.

[41] A. Mandelbaum and U. Yechiali. Optimal entering rules for a customer with wait option at an $M/G/1$ queue. *Management Science*, 29(2):174–187, 1983.

[42] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

[43] M. J. Meixell and M. Norbis. A review of the transportation mode choice and carrier selection literature. *The International Journal of Logistics Management*, 19(2):183–211, 2008.

[44] H. Min and G. Zhou. Supply chain modeling: Past, present and future. *Computers & Industrial Engineering*, 43(1):231–249, 2002.

[45] C. C. Moallemi, S. Kumar, and B. Van Roy. Approximate and data-driven dynamic programming for queueing networks. Submitted for publication, 2008.

[46] L. V. Montiel and J. E. Bickel. Generating a random collection of discrete joint probability distributions subject to partial information. *Methodology and Computing in Applied Probability*, 15(4):951–967, 2013.

[47] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

[48] E. Özkaya, P. Keskinocak, V. R. Joseph, and R. Weight. Estimating and benchmarking less-than-truckload market rates. *Transportation Research E*, 46(5):667–682, 2010.

[49] W. W. Qu, J. H. Bookbinder, and P. Iyogun. An integrated inventory-transportation system with modified periodic policy for multiple products. *European Journal of Operational Research*, 115(2):254–269, 1999.

[50] L. Rademacher. Approximating the centroid is hard. In *Proceedings of 23th Annual ACM Symposium of Computational Geometry*, pages 302–305, 2007.

[51] B. Q. Rieksts and J. A. Ventura. Optimal inventory policies with two modes of freight transportation. *European Journal of Operational Research*, 186(2):576–585, 2008.

[52] A. M. Sarmiento and R. Nagi. A review of integrated analysis of production-distribution systems. *IIE Transactions*, 31(11):1061–1074, 1999.

[53] S. Sindhuchao, H. E. Romeijn, E. Akçali, and R. Boondiskulchok. An integrated inventory-routing system for multi-item joint replenishment with limited vehicle capacity. *Journal of Global Optimization*, 32(1):93–118, 2005.

[54] S. R. Swenseth and M. R. Godfrey. Incorporating transportation costs into inventory replenishment decisions. *International Journal of Production Economics*, 77(2):113–130, 2002.

[55] H. Thorisson. Coupling methods in probability theory. *Scandinavian Journal of Statistics*, 22:159–182, 1995.

[56] S. Viswanathan and K. Mathur. Integrating routing and inventory decisions in one-warehouse multiretailer multiproduct distribution systems. *Management Science*, 43(3):294–312, 1997.

[57] W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212, 1999.

[58] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102, 2006.

[59] C. Winston. The demand for freight transportation: Models and applications. *Transportation Research A*, 17(6):419–427, 1983.

[60] J. Zan, J. J. Hasenbein, and D. P. Morton. Asymptotically optimal staffing of service systems with joint QoS constraints. *Queueing Systems*, 78(4):359–386, 2014.

[61] W. Zheng. Analysis of customer behavior under uncertainty on distribution of service time parameter in $M/M/1$ queueing system. In *12th International Conference on Service Systems and Service Management*, pages 1–5. IEEE, 2015.